ARTICLE IN PRESS

Information Sciences xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

Information Sciences

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

Evaluation of IPAQ questionnaires supported by formal concept analysis ☆

Radim Belohlavek [a,c,*], Erik Sigmund [b], Jiří Zacpal [c]

[a] *Department of Systems Science and Industrial Engineering, T. J. Watson School of Engineering and Applied Science, Binghamton University—SUNY, Binghamton, NY 13902, USA*

[b] *Center for Kinanthropology Research, Palacky University, Olomouc, tr. Miru 115, CZ-771 11 Olomouc, Czech Republic*

[c] *Dept. Computer Science, Palacky University, Olomouc, Tomkova 40, CZ-779 00 Olomouc, Czech Republic*

ARTICLE INFO

Article history:
Received 9 March 2009
Received in revised form 20 February 2010
Accepted 13 April 2010
Available online xxxx

Keywords:
Formal concept analysis
Questionnaire
Data visualization
Fuzzy logic

ABSTRACT

The paper presents a method for evaluation of questionnaires supported by formal concept analysis. Formal concept analysis provides an expert with a structured view on the data contained in the questionnaires. The method results from experiments with IPAQ (International Physical Activity Questionnaire). The structured view on the data provided by the method suggests various hypotheses which can later be tested. In addition, the structured view on data itself proved to be sufficiently informative to the expert. In addition to the method, the paper presents experiments with evaluation of IPAQ.

1. Introduction and problem setting

Questionnaires are used in many areas of human activity. They are used to reveal patterns of behavior and various kinds of dependencies among variables being surveyed. Descriptive statistics and statistical hypotheses testing are among the tools traditionally used for questionnaire evaluation [37]. A practical disadvantage of traditional statistical approaches is the need to formulate the hypotheses to be tested. Without any prior structured view on the data contained in the questionnaires, a formulation of relevant hypotheses is a difficult task. Another disadvantage of traditional statistical approaches is the limitation regarding what statistics can reveal about data and how statistical summaries can be understood by experts in the field of inquiry who are not experts in statistics.

This paper presents results on a particular way of evaluation of questionnaires that is supported by Formal Concept Analysis (FCA). The questionnaire used in this paper is the IPAQ (International Physical Activity Questionnaire). The paper is a continuation of previous studies concerning IPAQ [34,35]. IPAQ is a standardized international questionnaire [14]. At the beginning of our study, there was a need for an alternative means of evaluation of questionnaires formulated by experts from the Faculty of Physical Culture of the Palacky University, Olomouc, who are involved in a world-wide project of monitoring physical activities in today's population. The experts struggled with classic statistical techniques and were looking for alternative methods to evaluate questionnaires. Their principal concern was that statistical methods did not provide them with a global view of the data contained in the questionnaires from which relevant patterns could be seen. It turned out that the basic methods of Formal Concept Analysis (FCA), see [21], were useful for the experts. Namely, a concept lattice associated to

☆ Supported partly by Grant No. MSM 6198959214 "Mathematical Models and Structures", and by Grant No. 6198959221 "Physical activity and inactivity of inhabitants of the Czech Republic in the context of behavioral changes" of the MSM, and by Grant No. P103/10/1056 of the Czech Science Foundation.

* Corresponding author at: Department of Systems Science and Industrial Engineering, T.J. Watson School of Engineering and Applied Science, Binghamton University—SUNY, Binghamton, NY 13902, USA.

E-mail addresses: rbelohla@binghamton.edu (R. Belohlavek), erik.sigmund@upol.cz (E. Sigmund), jiri.zacpal@upol.cz (J. Zacpal).

0020-0255/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2010.04.011

Please cite this article in press as: R. Belohlavek et al., Evaluation of IPAQ questionnaires supported by formal concept analysis, Inform. Sci. (2010), doi:10.1016/j.ins.2010.04.011

the questionnaire data, and parts of the concept lattice, provide experts with an easy-to-understand hierarchically structured global view on the data. In terms of FCA, the basic idea is the following. Objects are the individual respondents (or groups of respondents) being surveyed in the questionnaires, attributes correspond to the variables being monitored by the questionnaires. The corresponding concept lattice, or its parts, reveal to the domain expert the groups of respondents categorized according to the attributes surveyed. The expert can see in the concept lattice various dependencies between attributes, the size of the respondents' groups, etc. Therefore, the concept lattice provides the expert with a first insight into the data. Such an insight is crucial. Very often, this insight is what the expert needs to see. Furthermore, based on this insight, the expert can pursue more detailed inquiries including those based on ordinary statistical techniques.

The present study focuses on considering groups of individuals as objects (objects in terms of FCA). The groups are defined by the expert in that they are based on sharing common attributes specified by the expert. The groups can be seen as aggregates. As a result, instead of sharing an attribute by the particular individuals from a specified group, we consider a relative frequency of the attribute within that group. In this way, one comes from data tables with yes/no attributes (i.e. ordinary formal contexts) to data tables with numbers from the unit interval $[0,1]$ which are interpreted as relative frequencies. We use particular fuzzy concept lattices as the concept lattices associated with such data tables. The concept lattices provide an expert with an aggregate hierarchical view on the data. The advantage of taking groups and the relative frequencies instead of individuals and original attributes is the conciseness of description provided by the resulting concept lattice which is what the experts asked for. The conciseness results from the loss of information which occurs as a result of aggregation. The disadvantage, as with other methods which involve aggregation and summarization, is loss of information.

The paper is organized as follows. Section 2 provides preliminaries from formal concept analysis. In Sections 3 and 4, we show how we represent the questionnaire data and how we obtain a fuzzy concept lattice from the data. Section 5 explains how the evaluation of the questionnaire is carried out. In Section 6, we summarize some of the results of the analysis of the IPAQ questionnaire with a data collected in a survey involving 4510 adults. Section 7 discusses methods related to our method, their relationship to it, and outlines some issues for future research.

2. Preliminaries from formal concept analysis

2.1. Basic aims

Formal Concept Analysis (FCA) is a method for data analysis with applications in various domains (see [21] for foundations and [11] for applications). The input to FCA consists of a data table describing a set $X = \{1,\ldots,n\}$ objects and a set $Y = \{1,\ldots,m\}$ of attributes. The table specifies attribute values of the objects. The main aim in FCA is to reveal from the data a hierarchically organized set of particular clusters, called formal concepts, and a small set of particular attribute dependencies, called attribute implications. FCA aims at formalizing and utilizing a traditional theory of concepts, in which a concept is understood as an entity consisting of its extent (collection of all objects to which the concept applies) and its intent (collection of all attributes to which are characteristic for the concept). For example, the extent of the concept *dog* consists of all dogs while its intent consists of attributes such as *barks*, *has tail*, *has four limbs*, etc. The information extracted from data in FCA is well-comprehensible by users because FCA works with notions which humans are used to reason with in the ordinary life. This is an important feature of FCA which makes it appealing for users.

2.2. Data with binary attributes and conceptual scaling

In the basic setting of FCA, attributes are assumed to be binary, i.e. a given object either has or does not have a given attribute. A data table with binary attributes is represented by a triplet $\langle X,Y,I\rangle$, called a *formal context*, which consists of the above-mentioned sets X and Y of objects and attributes, and a binary relation I between X and Y (incidence relation, to-have relation). Thus, $I \subseteq X \times Y$, $\langle x,y\rangle \in I$ indicates that object x has attribute y, $\langle x,y\rangle \notin I$ indicates that x does not have y. Objects $x \in X$ correspond to table rows, attributes $y \in Y$ correspond to table columns, and I is represented by *0*s and *1*s in the table entries (i.e. *1* indicates the corresponding object does have the corresponding attribute). A formal concept of $\langle X,Y,I\rangle$ is any pair $\langle A,B\rangle$ of sets $A \subseteq X$ (*extent*) and $B \subseteq Y$ (*intent*) such that B is just the set of attributes shared by all objects from A, and A is the set of all objects sharing all attributes from B. In symbols, this can be written as $A^{\uparrow} = B$ and $B^{\downarrow} = A$, where

$$A^{\uparrow} = \{y \in Y |\text{ for each } x \in A : \langle x,y\rangle \in I\},$$
$$B^{\downarrow} = \{x \in X |\text{ for each } y \in B : \langle x,y\rangle \in I\}.$$

The thus introduced arrow operators form a Galois connection between X and Y and play an important role in FCA. The set of all formal concepts of $\langle X,Y,I\rangle$ is denoted by $\mathcal{B}(X,Y,I)$. That is,

$$\mathcal{B}(X,Y,I) = \{\langle A,B\rangle | A^{\uparrow} = B, B^{\downarrow} = A\}.$$

Under partial order $\leqslant$ defined for $\langle A_1,B_1\rangle, \langle A_2,B_2\rangle \in \mathcal{B}(X,Y,I)$ by

$$\langle A_1,B_1\rangle \leqslant \langle A_2,B_2\rangle \text{ iff } A_1 \subseteq A_2 (\text{iff } B_2 \subseteq B_1)$$

Please cite this article in press as: R. Belohlavek et al., Evaluation of IPAQ questionnaires supported by formal concept analysis, Inform. Sci. (2010), doi:10.1016/j.ins.2010.04.011

$\mathcal{B}(X,Y,I)$ happens to be a complete lattice, so-called *concept lattice* associated to $\langle X,Y,I\rangle$ [21,38]. Efficient algorithms for computing $\mathcal{B}(X,Y,I)$ exist; a good overview is provided by [28].

FCA can handle *many-valued attributes*, i.e. more general attributes such as ordinal or nominal, by means of so-called *conceptual scaling* [21]. A data table with many-valued attributes is called a *many-valued formal context* in FCA. In the following, we use a simple scaling in which a set Z of many-valued attributes describing objects of a set X is transformed to an ordinary formal context $\langle X,Y,I\rangle$ as follows. For every attribute $z\in Z$, we partition the set rng (z) of all values of z into disjoint subsets. For every such subset $V\subseteq$ rng (z), one creates and adds to Y a new attribute $y_{z,V}$. The new incidence relation I between X and the thus created set Y of attributes is defined as follows: $\langle x,y_{z,V}\rangle\in I$ if and only if the value of z on x belongs to V. For example, if z is *age* with rng $(z)=[0,1,\ldots,100]$, sets V may be the intervals $V_{[0,9]}=[0,1,\ldots,9],\ldots,V_{[90,100]}=[90,91,\ldots,100]$, and we put $\langle x,y_{z,V_{[0,9]}}\rangle\in I$ if and only if the age of person x is in $[0,1,\ldots,9]$.

2.3. Data with fuzzy attributes

Several extensions of FCA have been proposed for data which contain fuzzy attributes, i.e. attributes such as *good performance* which apply to objects to degrees. We use the approach described in e.g. [2,31] and refer to [3,5,6] for overview of other approaches. A data table with fuzzy attributes can be represented by a triplet $\langle X,Y,I\rangle$, called a formal fuzzy context, where X and Y is the set of objects and attributes, as above, and $I\in\mathbf{L}^{X\times Y}$ is a binary fuzzy relation between X and Y assigning to each object $x\in X$ and each attribute $y\in Y$ the degree $I(x,y)\in L$ to which x has y. Here, $\mathbf{L}=\langle L,\wedge,\vee,\otimes,\rightarrow,0,1\rangle$ is a structure of truth degrees consisting of a set L of truth degrees and operations $\wedge$, $\vee$, $\otimes$, $\rightarrow$ (truth functions of logical connectives) for manipulation of truth degrees. In particular, $\rightarrow$ is the truth function of implication. L contains 0 (full falsity, smallest truth degree) and 1 (full truth, largest truth degree) and possibly intermediate truth degrees. We assume that $\mathbf{L}$ forms a complete residuated lattice and refer to [2,22,25] for more information about fuzzy sets and fuzzy logic. The above notions generalize to the case of fuzzy attributes as follows. For $A\in\mathbf{L}^X$,$B\in\mathbf{L}^Y$ (i.e. A is a fuzzy set of objects, B is a fuzzy set of attributes), we define fuzzy sets $A^{\uparrow}\in\mathbf{L}^Y$ (fuzzy set of attributes) and $B^{\downarrow}\in\mathbf{L}^X$ (fuzzy set of objects) by

$$A^{\uparrow}(y)=\bigwedge\nolimits_{x\in X}(A(x)\rightarrow I(x,y)),\qquad(1)$$

$$B^{\downarrow}(x)=\bigwedge\nolimits_{y\in Y}(B(y)\rightarrow I(x,y)).\qquad(2)$$

Described verbally, $A^{\uparrow}$ is the fuzzy set of all attributes from Y shared by all objects from A (and similarly for $B^{\downarrow}$). A formal concept of $\langle X,Y,I\rangle$ is any pair $\langle A,B\rangle$ of $A\in\mathbf{L}^X$ and $B\in\mathbf{L}^Y$ satisfying $A^{\uparrow}=B$ and $B^{\downarrow}=A$. The collection $\mathcal{B}(X,Y,I)=\{\langle A,B\rangle|A^{\uparrow}=B,B^{\downarrow}=A\}$, i.e. the collection of all formal concepts of $\langle X,Y,I\rangle$, can be equipped with a partial order $\leqslant$ defined by $\langle A_1,B_1\rangle\leqslant\langle A_2,B_2\rangle$ iff $A_1\subseteq A_2$ (iff $B_2\subseteq B_1$). Here, $A_1\subseteq A_2$ means that $A_1(x)\leqslant A_2(x)$ for each $x\in X$ (similarly for $B_2\subseteq B_1$). $\mathcal{B}(X,Y,I)$ equipped with $\leqslant$ is a complete lattice, called a (fuzzy) concept lattice associated to $\langle X,Y,I\rangle$ [2,31]. Various parts of $\mathcal{B}(X,Y,I)$ have been subject to investigation. In particular, we use the so-called *crisply generated formal concepts* of $\langle X,Y,I\rangle$ [4], i.e. formal concepts $\langle A,B\rangle\in\mathcal{B}(X,Y,I)$ generated by a crisp set C of objects. Note that $C\in L^X$ is called crisp if $C(x)=0$ or $C(x)=1$ for each object $x\in X$. The set of all crisply generated formal concepts of $\langle X,Y,I\rangle$ is called a crisply generated fuzzy concept lattice and is denoted by $\mathcal{B}_c(X,Y,I)$. We thus have

$$\mathcal{B}_c(X,Y,I)=\{\langle A,B\rangle\in\mathcal{B}(X,Y,I)|A=C^{\downarrow}\text{ for some crisp }C\in L^X\}.$$

Note that crisply generated fuzzy concept lattices are isomorphic to one-sided fuzzy concept lattices defined in [26,27] as well as to the fuzzy concept lattices defined in [9], cf. also [8,39].

3. The questionnaire data and its representation in FCA

Each IPAQ questionnaire consists of questions to be answered by respondents by selecting an answer from a list of possible answers. From the point of view of FCA, we can consider the set of respondents as the set of objects and the set of single questions as the set of attributes. The questions need not be yes/no questions. Rather, some questions like those concerning age and education are many valued. Correspondingly, a completed questionnaire can be represented by a many-valued formal context. Such a context can be transformed to an ordinary formal context $\langle X,Y,I\rangle$ via conceptual scaling, see Section 2. That is, X is the set of individual respondents, Y represents answers to the questions, and the binary relation $I\subseteq X\times Y$ between X and Y represents how respondents answered the questions. The attributes in Y are defined as follows. Let q be a question whose set of possible answers is rng (q). We partition rng (q) into disjoint subsets $V_1,\ldots,V_{n_q}\subseteq\text{rng}(q)$ and for every V_i we add to Y an attribute y_i. Now, we put $\langle x,y_i\rangle\in I$ if the answer of respondent x to question q is an element of V_i. Y then consists of all attributes y_i which result this way. Thus, for the question inquiring whether a respondent has a job or not, we add two attributes, JOByes and JOBno. If respondent x answers "yes" to the question, we have $\langle x,\text{JOByes}\rangle\in I$ and $\langle x,\text{JOBno}\rangle\notin I$. For the question regarding respondent's age, we add attributes AGE15–24,…,AGE45–54, AGE55–65. Note that as usual with data binarization, one needs to consult with the expert to define meaningful binary attributes.

Typically, a formal context that results from a collection of questionnaires contains many objects and a manageable number of attributes. Nevertheless, the corresponding concept lattice is too large for an expert to comprehend. In addition, the expert might not be interested in the formal concepts from this concept lattice. Rather, the expert might want to consider

Please cite this article in press as: R. Belohlavek et al., Evaluation of IPAQ questionnaires supported by formal concept analysis, Inform. Sci. (2010), doi:10.1016/j.ins.2010.04.011

Table 1
The original formal context from 19 survey respondents.

	GENDERmale	GENDERfemale	JOByes	JOBno	AGE15–24	AGE25–34	AGE35–44	AGE45–54	AGE55–65	BMIunder	BMInormal	BMIover	BMIobesity	PAlow	PAmoderate	PAhigh
1	1	0	1	0	0	0	0	1	0	0	0	1	0	1	0	0
2	1	0	1	0	1	0	0	0	0	0	1	0	0	1	0	0
3	1	0	1	0	0	0	0	1	0	0	1	0	0	1	0	0
4	1	0	1	0	0	0	0	1	0	0	1	0	0	1	0	0
5	1	0	1	0	0	0	0	1	0	0	1	0	0	1	0	0
6	1	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0
7	1	0	0	1	1	0	0	0	0	0	1	0	0	1	0	0
8	1	0	0	1	0	0	0	1	0	0	1	0	0	1	0	0
9	1	0	0	1	0	0	0	1	0	0	1	0	0	1	0	0
10	0	1	1	0	1	0	0	0	0	1	0	0	0	1	0	0
11	0	1	1	0	0	0	1	0	0	0	1	0	0	1	0	0
12	0	1	1	0	1	0	0	0	0	0	1	0	0	1	0	0
13	0	1	1	0	1	0	0	0	0	1	0	0	0	1	0	0
14	0	1	1	0	1	0	0	0	0	1	0	0	0	1	0	0
15	0	1	1	0	0	1	0	0	0	1	0	0	0	1	0	0
16	0	1	0	1	0	0	0	0	1	1	0	0	0	1	0	0
17	0	1	0	1	1	0	0	0	0	1	0	0	0	1	0	0
18	0	1	0	1	1	0	0	0	0	1	0	0	0	1	0	0
19	0	1	0	1	1	0	0	0	0	1	0	0	0	1	0	0

Table describing 19 objects (rows) and 16 attributes (columns). Attributes BMIunder,...,BMIobesity are derived from the Body Mass Index (BMI). Attributes PAlow, PAmoderate, and PAhigh are based on the level of Physical Activity (PA).

Please cite this article in press as: R. Belohlavek et al., Evaluation of IPAQ questionnaires supported by formal concept analysis, Inform. Sci. (2010), doi:10.1016/j.ins.2010.04.011

Table 2
Formal fuzzy context derived from the formal context from Table 1.

	AGE15–24	AGE25–34	AGE35–44	AGE45–54	AGE55–65	BMIunder	BMInormal	BMIover	BMIobesity	PAlow	PAmoderate	PAhigh
Fno	0.75	0	0	0	0.25	1	0	0	0	1	0	0
Fyes	0.67	0.17	0.17	0	0	0.67	0.33	0	0	1	0	0
Mno	0.25	0	0.25	0.5	0	0.25	0.75	0	0	1	0	0
Myes	0.20	0	0	0.80	0	0	0.80	0.20	0	1	0	0

aggregates of the individual respondents (namely, groups of respondents) as objects in the formal context, such as "men", "women who have a job", etc. Correspondingly, one might want to define the aggregates by having the same attributes from a set S of attributes specified by an expert, such as those regarding age, gender, etc., with S being a subset of the set Y of all attributes. We call the attributes from S *characteristic attributes*. The aggregates we consider are equivalence classes of individual respondents.

Definition 1. Let $S \subseteq Y$ be a set of (characteristic) attributes. For respondents $x_1, x_2 \in X$, put

$$x_1 \equiv_S x_2 \quad \text{if and only if } \{x_1\}^{\uparrow} \cap S = \{x_2\}^{\uparrow} \cap S,$$

i.e.

$$x_1 \equiv_S x_2 \quad \text{if and only if } \{y \in S | \langle x_1, y\rangle \in I\} = \{y \in S | \langle x_2, y\rangle \in I\}.$$

In words, $x_1 \equiv_S x_2$ if and only if respondents x_1 and x_2 have the same attributes from S, i.e. x_1 and x_2 are indistinguishable by the attributes from S. Clearly, $\equiv_S$ is an equivalence relation on X. We call the classes $[x]_{\equiv_S}$ of $\equiv_S$ *aggregate objects*. Note that $[x]_{\equiv_S}$ is the set of all objects equivalent with x, i.e. $[x]_{\equiv_S} = \{x' \in X | x \equiv_S x'\}$. Furthermore, denote

- by X_S the set of all classes of $\equiv_S$, i.e. $X_S = X/\equiv_S$,
- by Y_S the set of those attributes from Y not included in S, i.e. $Y_S = Y - S$.

Now, for each class $[x]_{\equiv_S}$ from X_S and each attribute $y \in Y_S$, we consider the relative frequency of y in class $[x]_{\equiv_S}$ and denote it by $I_S([x]_{\equiv_S}, y)$. That is, we put

$$I_S([x]_{\equiv_S}, y) = \frac{|\{x_1 \in [x]_{\equiv_S} : \langle x_1, y\rangle \in I\}|}{|[x]_{\equiv_S}|}.$$

We consider I_S a fuzzy relation. Namely, we consider the fuzzy concept lattice $\mathcal{B}(X_S, Y_S, I_S)$ associated to $\langle X_S, Y_S, I_S\rangle$ and, in particular, its subset $\mathcal{B}_c(X_S, Y_S, I_S)$—the set of crisply generated fuzzy concepts, see Section 2. For simplicity, we round the degrees assigned by I_S to those from the scale $\{0, 0.01, \ldots, 0.99, 1\}$.

Remark 1. An important remark is in order. Interpreting degrees of membership of a fuzzy relation by relative frequencies is not typical. It may even seem inappropriate because degrees in fuzzy logic are typically degrees associated with graded (fuzzy) collections and relationships. However, as we will see, the way we use the degrees with frequential interpretation is correct from the semantical point of view. In fact, the only thing that matters in our manipulation with the degrees is a comparison of the degrees, i.e. a comparison of relative frequencies.

In addition, an expert might want not to include aggregate objects $[x]_{\equiv_S}$ which contain less than m objects with m being a certain threshold prescribed by an expert (this requirement was suggested by the domain expert).

Example 1. Consider the following illustrative example. Let the ordinary formal context be given by Table 1. Consider S = {GENDERmale, GENDERfemale, JOByes, JOBno} as the set of characteristic attributes. Using the above-described transformation, we obtain a formal fuzzy context with four aggregate objects Fno (women who do not have a job), Fyes (women who have a job), Mno (men who do not have a job), Myes (men who have a job), depicted in Table 2. Note that in this particular example, the same set of aggregate objects results from S = {GENDERmale, JOByes} or from S = {GENDERfemale, JOBno} since the value of GENDERmale uniquely determines the value of GENDERfemale and *vice versa*, and the same for JOByes and JOBno.

4. Concept lattices from the derived contexts with relative frequencies

Let $\langle X, Y, I\rangle$ be a formal context representing a questionnaire data as described above and let S be a set of characteristic attributes selected by a user. As described in the previous section, $\langle X, Y, I\rangle$ can be transformed to a formal fuzzy context $\langle X_S, Y_S, I_S\rangle$ consisting of a set X_S of aggregate objects, a set Y_S of their attributes, and a fuzzy relation I_S with $I_S([x]_{\equiv_S}, y)$

representing the relative frequency of attribute y in class $[x]_{\equiv_S}$. We use the crisply generated fuzzy concept lattice $\mathcal{B}_c(X_S, Y_S, I_S)$ for displaying the information contained in the questionnaires. That is, in our particular setting, we have

$$\mathcal{B}_c(X_S, Y_S, I_S) = \{\langle A, B\rangle | A^{\uparrow} = B, \quad B^{\downarrow} = A, \quad \text{and} \quad A = C^{\uparrow\downarrow} \quad \text{for some } C \in \{0,1\}^{X_S}\},$$

where the operators are defined by

$$A^{\uparrow}(y) = \bigwedge_{[x]_{\equiv_S} \in X_S} (A([x]_{\equiv_S}) \to I_S([x]_{\equiv_S}, y)),$$

$$B^{\downarrow}([x]_{\equiv_S}) = \bigwedge_{y \in Y_S} (B(y) \to I_S([x]_{\equiv_S}, y)).$$

Remark 2. Formal concepts $\langle A,B\rangle$ from $\mathcal{B}_c(X_S, Y_S, I_S)$ can be interpreted as follows. A is a collection of aggregate objects (i.e. groups of individual respondents) and B is a fuzzy set such that every aggregate object (group of individual respondents) from A has attribute y with relative frequency at least $B(y)$.

We use such an interpretation for the sake of simplicity. Namely, strictly speaking, A is a fuzzy set, too. Therefore, we should say that $\{[x]_{\equiv_S} \in X_S | A([x]_{\equiv_S}) = 1\}$ is the set of aggregate objects such that every $[x]_{\equiv_S} \in A$ has attribute y with relative frequency at least $B(y)$. For aggregate objects $[x]_{\equiv_S}$ with $A([x]_{\equiv_S}) < 1$, high values $A([x]_{\equiv_S})$ indicate that the relative frequencies of ys are close to $B(y)$.

5. Evaluation of data collected from questionnaires

A fuzzy concept lattice $\mathcal{B}_c(X_S, Y_S, I_S)$ proved to be a suitable tool for domain expert exploration of the data collected from the questionnaires. The expert acquires his knowledge by browsing the line diagram of the concept lattice, or its parts. In particular, the expert explores the formal concepts (they represent groups of respondents) and the information associated to the formal concepts, such as the objects and aggregate objects represented by the extent A of a formal concept $\langle A,B\rangle$, the number of such objects covered by a formal concept, as well as the attributes and their relative frequencies represented by the intent B of $\langle A,B\rangle$. Formal concepts are represented by the nodes of the line diagram. The lines of the diagram represent the subconcept–superconcept hierarchy of the respondents' groups. Exploring the hierarchy helps the expert see various dependencies between attributes which exist in the data.

A frequent expert request is to look for properties which are common to groups of respondents (i.e. aggregate objects which we represent by characteristic attributes). For this purpose, one can efficiently retrieve the smallest formal concept, which includes all requested aggregate objects (i.e. the smallest formal concept $\langle A,B\rangle$ such that $A([x]_{\equiv_S}) = 1$ for every aggregate object $[x]_{\equiv_S}$ specified by the expert). The expert also wants to know how many and which respondents are contained in such a concept. This number is the sum of the numbers of objects contained in the aggregate objects, i.e. equal to

$$\sum_{A([x]_{\equiv_S})=1} |[x]_{\equiv_S}|.$$

Another feature frequently used by domain experts was an easy-to-understand characterization of the groups of respondents corresponding to formal concepts. Such characterization is due to a simple definition of aggregate objects based on characteristic attributes, cf. Definition 1. For example, with the set S of characteristic objects as in Example 1, if the extent of a given formal concept contains an aggregate object $[x_1]_{\equiv_S}$, for which x_1 has the characteristic attributes GENDERfemale and JOBno, and an aggregate object $[x_2]_{\equiv_S}$, for which x_2 has the characteristic attributes GENDERmale and JOByes, the expert can easily see that this formal concept represents a group of respondents that includes women who do not have a job as well as men who do have a job.

Another type of information is contained in an intent B of a formal concept $\langle A,B\rangle$. The degree $B(y)$ expresses the percentage of objects that are present in the extent A (in degree 1) and have attribute y. More precisely, for every aggregate object $[x]_{\equiv_S}$ of $\langle A,B\rangle$ (i.e. such that $A([x]_{\equiv_S}) = 1$), at least $100 \cdot B(y)\%$ of objects from $[x]_{\equiv_S}$ have attribute y. Hence, also at least $100 \cdot B(y)\%$ of objects that belong to the union of the aggregate objects $[x]_{\equiv_S}$ of $\langle A,B\rangle$ have attribute y. For example, consider the formal concept $\langle A,B\rangle$ whose extent is described above, i.e. the formal concept contains an aggregate object $[x_1]_{\equiv_S}$, for which x_1 has the characteristic attributes GENDERfemale and JOBno, and an aggregate object $[x_2]_{\equiv_S}$, for which x_2 has the characteristic attributes GENDERmale and JOByes. If attribute SMOKINGno belongs to B to degree *0.42* and attribute DOGyes to degree *0.65*, this means that at least 42% women without a job and at least 42% of working men do not smoke. Hence, also at least 42% respondents who are women without a job or working men do not smoke. Similarly, at least 65% respondents who are women without a job or working men have a dog. Both high and low values of $B(y)$ are interesting for the expert.

By traversing the concept lattice downward, we can examine common attributes of different subsets of aggregate objects. A comparison of formal concepts created from aggregate objects is also interesting for the expert. For example, if the expert is about to examine the influence of physical activity on population, then comparing a formal concept that contains the aggregate object whose respondents have attributes GENDERmale and PAlow, and the aggregate object whose respondents have attributes GENDERfemale and PAlow, with a formal concept which contains the aggregate object whose respondents

Please cite this article in press as: R. Belohlavek et al., Evaluation of IPAQ questionnaires supported by formal concept analysis, Inform. Sci. (2010), doi:10.1016/j.ins.2010.04.011

have attributes GENDERmale and PAhigh, and the aggregate object whose respondents have attributes GENDERfemale and PAhigh is of immediate interest to the expert. For instance, the expert is able to see from the intents of these concepts in which attributes the group of females with a low level of physical activity is different from the group of females with a high level of physical activity. In addition, by looking at neighboring formal concepts and examining their groups of respondents and attributes, one can see various dependencies between attributes. The next section shows in detail some results of these analyses.

6. IPAQ, results of evaluation, and user experience

This section presents a closer look at the IPAQ questionnaire and the problems regarding its evaluation that the domain experts attempted to solve. It also demonstrates by means of concrete examples how the method explained in the previous sections was used by the experts.

IPAQ: The International Physical Activity Questionnaire (IPAQ) is a world-wide questionnaire used for monitoring Physical Activity (PA) among 18–65-year-old adults in diverse (international) populations [14,40]). The short self-administered format of IPAQ measures the frequency and duration of walking, moderate-intensity and vigorous physical activity, for leisure, transportation, and occupational purposes, and that of sitting during past week. Every IPAQ participant answers questions regarding the time spent performing physical activities, various personal data (such as gender, age, body weight and height), and lifestyle characteristics (such as education, time spent at work, type of housing—own house × apartment, family status—living alone × in family × in family with children, smoking—yes × no, ownership of car, dog, bike, and weekend house, and participation in an organized physical activity).

Reliability and validity of IPAQ were evaluated with over 2500 adults from 12 countries [14]. One-week test–retest reliability of the short self-administered format of IPAQ is very good (Spearman $r = 0.70–0.97$). The criterion validity for the IPAQ total minutes per week was acceptable as measured against accelerometer total counts (Spearman $r = 0.23$) and for the average correct classification of respondents accumulating $\geqslant 150$ min per week of physical activity (Spearman $r = 0.74$) [14].

The most important result from IPAQ is the finding of how many adults (% of the surveyed sample) meet the category "low", "moderate" or "high" levels of PA in relation to health-related recommendations [1]. A complex and transparent description of groups of adults with different PA levels is sought. This description should include environmental, social, and possibly other factors determining the lifestyle of the participants [15,32,33]).

Questionnaire data used: For the analysis of IPAQ questionnaire data we used the data collected in a survey regarding the physical activity of population of the Czech republic conducted April and May of 2006. 7934 adults (53% female, 47% male) were selected to participate in the survey. The participants were selected by randomly generating postal addresses from all regions of the Czech Republic. Every participant received an envelope containing a letter with a basic information about the survey and its purpose, the IPAQ questionnaire with instructions, and an empty envelope with a stamp and with the address where to send the completed questionnaire. 5236 completed questionnaires were received from the 7934 selected respondents (66%). Incorrectly completed questionnaires and questionnaires with missing data were excluded, resulting thus into 4510 questionnaires completed by the respondents (2393 female, 2117 male, age 18–65 years). The 4510 respondents were subject to the analysis and evaluation as described below.

Results of IPAQ evaluation: Simple and multiple correlation analyses carried out by the domain experts did not show deeper associations between PA levels and personal data and lifestyle characteristics of participating adults. Therefore, it was not possible to conclude that smokers or obese adults are less physically active than non-smokers or non-obese adults, due to very low correlations between PA level and smoking or body weight. This is the main reason why the domain experts were looking for a different data analysis tool that would provide complex and transparent characterization of groups of adults with respect to their PA level and lifestyle characteristics. Formal concept analysis and the method proposed in this paper proved useful for this purpose.

The following questions are typical questions asked by the domain experts and those that the proposed method helped answer. Are there any differences between groups of females with low and high levels of PA in relation to the guidelines of healthy lifestyle? Are there any similarities in personal data or lifestyle characteristics in these two groups of females? Are any of these characteristics more typical for a group of females with high levels of PA than for a group of females with low PA levels?

The entire questionnaire data was transformed into an ordinary formal context $\langle X, Y, I\rangle$ that contained 4510 objects and 47 attributes. Appendix describes the 47 attributes as well as how they were derived from the original IPAQ data. Recall from Section 3 that objects represent respondents and that attributes represent answers to questions. As the next step, the expert specified the characteristic attributes, i.e. defined the aggregate objects (see Section 3). The set S of characteristic attributes contained attributes GENDERmale, GENDERfemale, PAlow, PAmoderate, and PAhigh, i.e.

$$S = \{\text{GENDERmale}, \ \text{GENDERfemale}, \ \text{PAlow}, \ \text{PAmoderate}, \ \text{PAhigh}\}.$$

As explained in Section 3, the characteristic attributes produce a formal fuzzy context $\langle X_S, Y_S, I_S\rangle$ with 6 aggregate objects and 42 attributes. The aggregate objects are denoted by MLo (men with low PA), MMo (men with moderate PA), MHi (men with high PA), FLo (women with low PA), FMo (women with moderate PA), FHi (women with high PA). The resulting fuzzy concept lattice $\mathcal{B}_c(X_S, Y_S, I_S)$ consists of 54 crisply generated formal concepts.

Please cite this article in press as: R. Belohlavek et al., Evaluation of IPAQ questionnaires supported by formal concept analysis, Inform. Sci. (2010), doi:10.1016/j.ins.2010.04.011

An exploration of this concept lattice helps answer the questions formulated above. For the purpose of illustration, we include the following concrete example. If the expert is interested in data related to women, he will work with the part of the concept lattice depicted in Fig. 1. This part contains formal concepts with extents containing aggregate objects FLo, FMo, or FHi (i.e. the respective concept extents contain the aggregate objects FLo, FMo, and FHi to degree 1). For example, as one can see from Fig. 1, there is a high number of women with at least moderate PA level. This result is strongly influenced by everyday walking (50% of women spend more than 65 min per day walking). Ownership of a car or a dog, living in house, living alone, or smoking and having a job are the attributes in which the groups corresponding to FLo and FHi do not differ much. On the other hand, participation in an organized PA more than 3 times per week and high walking activity (more than 1 hour per day) are the attributes in which the groups corresponding to FLo and FHi significantly differ. This and other information can be concluded from Table 3 which shows the intents of formal concepts corresponding to FHi (women with high level of PA) and FLo (women with low level of PA). These results correspond to findings of [36]. Surprisingly, the values of AGE15–24 for FHi and FLo indicate that physical activity of younger people is lower than that of older people, which contradicts the results obtained in [19].

Another type of problem formulated by the expert was to see which features are shared by certain aggregate objects, i.e. by certain groups of respondents. For example, what is common to women and men with high levels of physical activity as compared to women and men with low levels of physical activity? Such questions can be answered by examining the intent of the corresponding formal concept, i.e. the formal concept whose extent contains aggregate objects FHi and MHi (women and men with high level of PA) and a formal concept whose extent contains aggregate objects FLo and MLo (women and men with low level of PA), see Table 4. The intents show the main differences and similarities between women and men with high levels of PA and women and men with low levels of PA. High levels of PA of women and men are strongly associated with walking more than 60 min per day (Walking high) and cycling. These women and men had also a health-favorable BMI. Women and men with low levels of physical activity often walked less than 30 min per day and are more overweight than women and men with high level of physical activity, see Table 4.

Other important findings revealed in this analysis include the following somewhat surprising facts. People without a job have approximately the same level of PA as people with a job. Smoking seems to have no influence on the level of PA of people. For both of these findings, more information is needed for explanation. For example, are smokers physically active because they walk to smoke?

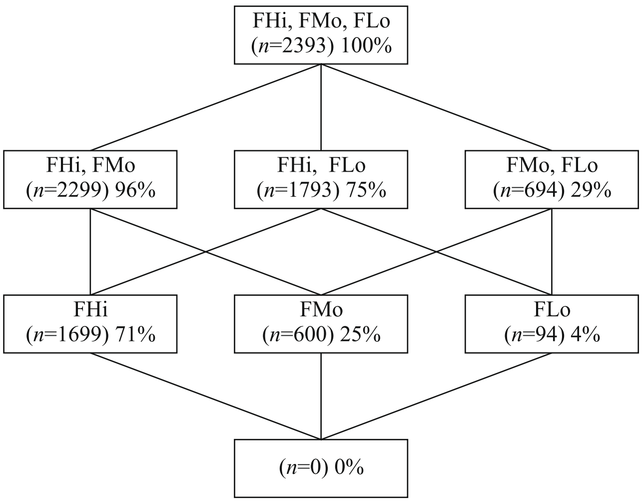

Fig. 1. Hierarchical diagram of formal concepts representing groups of women differentiated according to level of physical activity. n denotes the number of respondents (women) that belong to the group of respondents represented by the formal concepts. x% indicates that $x = (n/N)^{*}$ 100% of women respondents belong to the group represented by the formal concept (N is the number of all women respondents participating in the questionnaire).

Table 3
Differences between groups of women with high and low level of Physical Activity (PA).

	Lives in house	Lives alone	Owns car	Smokes	Has dog	Organized PA	Owns weekend house	BMIunder	BMInormal	Walking low	Walking middle	Job yes	Age 15–24
FHi	0.51	0.08	0.71	0.21	0.34	0.15	0.30	0.09	0.58	0.01	0.02	0.68	0.40
FLo	0.51	0.11	0.69	0.24	0.38	0.00	0.23	0.05	0.55	0.40	0.36	0.70	0.52

Please cite this article in press as: R. Belohlavek et al., Evaluation of IPAQ questionnaires supported by formal concept analysis, Inform. Sci. (2010), doi:10.1016/j.ins.2010.04.011

Table 4
Attributes shared by women and men with a high level of physical activity (row FHi,MHi) vs. women and men with a low level of physical activity (row FLo,MLo).

	Owns bicycle	Owns car	Job yes	Smokes	Has dog	Lives in house	Walking high	Walking low	BMIunder	BMInormal	BMIover	BMIobesity
FHi,MHi	0.86	0.71	0.68	0.21	0.34	0.51	0.52	0.01	0.09	0.58	0.09	0.08
FLo,MLo	0.75	0.69	0.70	0.24	0.30	0.51	0.24	0.40	0.01	0.33	0.32	0.08

The findings illustrated above are considered of major health-policy concern because of a documented increase of obesity in young people in today's population [24]. Growing obesity due to a physical inactivity represents one of the most costly diseases, in terms of treatment costs, in developed countries [13] including the Czech Republic [19]. Therefore, it is considered very important to search for determinants of low and high levels of physical activity. According to the results of the study reported in this paper as well as the results from [20], the Czech Republic can still be considered a country of walkers and cyclists. Therefore, a practical recommendation for a governmental policy, based on the results of this study and [20], is to support a development of urban and suburban environment which encourages both walking and cycling.

User experience: Let us summarize the expert experience with using the proposed method based on formal concept analysis.

– The proposed method leads to interesting findings in a short period of time. In particular, the expert did not have much prior knowledge of what the information hidden in the data might be and the method proved to be useful for data exploration given these conditions. Namely, the expert was able to perform quickly a rough exploratory analysis using the method to reveal basic patterns and possible hypotheses from the data.
– The expert found the method of more use than the methods which he was using before (regression and correlation analysis). Namely, the previously used methods are better suitable for a detailed quantitative analysis of known hypotheses regarding data dependencies rather than an exploratory analysis. The presented method met better the expert's needs: rough exploratory analysis with little knowledge about the data.
– The proposed method provides the expert with a considerable level of interactivity. The expert termed the proposed method "dynamic", compared to the statistical methods used before (regression and correlation analysis) which he considered "static".
– The expert can employ his partial knowledge or expectations regarding the questionnaire data in a natural way in the proposed method by selecting characteristic attributes and visually exploring the parts of the concept lattice which are of interest.
– The hierarchical diagram of a concept lattice provides the expert with both qualitative and quantitative information regarding the questionnaire data.
– Browsing the hierarchical diagram of a concept lattice enables the expert to generalize and specialize his hypotheses regarding the questionnaire data.

7. Related methods, further topics, and future work

In this section, we mention some related approaches to the analysis of questionnaires and outline some directions for future research.

An interesting method for evaluation of questionnaires based on formal concept analysis was described in [16]. The method utilizes a concept lattice and, in particular, attribute implications. Recall that an attribute implication is an expression $\{y_1, \ldots, y_m\} \Rightarrow \{z_1, \ldots, z_n\}$. An attribute implication holds true in a data (a formal context) if for every object x, the following is true: if x has all the attributes $y_1, \ldots, y_m$ then x has also all the attributes $z_1, \ldots, z_n$. In [16], the authors extract from the questionnaire data a small subset of the set of all attribute implications, called a basis [21]. The basis carries the information about all attribute implications true in the data. Namely, any attribute implication true in the data if it semantically follows from the basis. The basis is then used in [16] to analyze relationships between the attributes, i.e. between the answers to the questionnaire. The difference of the method described in [16] from our method is twofold. First, we directly utilize the concept lattice associated to the input data, while [16] does not make a direct use of this lattice. Second, we use a particular data transformation (aggregation) to obtain data in which the objects are relevant groups of respondents rather than particular respondents. Due to the transformation, we obtain a formal fuzzy context and use fuzzy concept lattices rather than ordinary concept lattices. For a future research, one direction is to apply the method from [16] to the IPAQ questionnaire. One possibility here is to extract the attribute implications from the ordinary formal context (data with binary attributes) obtained from the questionnaires (i.e. before aggregation). Another possibility is to extract the attribute implications from the formal fuzzy context (data with fuzzy attributes) which results from the aggregation. To do this, one can use the theory and algorithms available for fuzzy attribute implications, see [7] for an overview.

Another related method was presented in [10]. The authors use attribute implications equipped with a statistically motivated semantics which resembles GUHA-quantifiers [23]. That is to say, the authors consider an approximate validity of an attribute implication rather than an absolute validity, meaning that an attribute implication is considered interesting even if the are a few counterexamples to it in the data. Again, the method differs from the method proposed in this paper in its aims. Namely, the output of the method presented to a user consists of attribute implications, rather than parts of a concept lattice. Moreover, the method does not use any transformation of the input data. Besides the idea to apply the method from [10] to the formal context obtained directly from the questionnaires, an possible idea for the future research is to develop theory and algorithms of approximate validity of attribute implications for data with fuzzy attributes in the spirit of the GUHA method [23]. Note that the GUHA method is a method of exploratory data analysis, based on logical and statistical reasoning, with advanced theoretical and computational foundations (the foundations are comprehensively covered in [23]). While GUHA takes data with binary attributes at the input, the topic for future research mentioned above is to generalize the methods in order to work directly for data with fuzzy attributes.

Another method related to the approach presented in this paper is COBWEB [17] and, in general, the methods of conceptual clustering developed within machine learning [29]. Conceptual clustering methods accept objects described in terms of their attributes and produce a classification scheme from the object descriptions. COBWEB itself is an incremental method of conceptual clustering, i.e. it processes the objects one by one and updates correspondingly the system of clusters being created. Different order of processing of the objects may result in different systems of clusters. Moreover, the system of clusters produced forms a tree. This is one difference from the methods of formal concept analysis from a general viewpoint. Another general difference is that COBWEB favors clusters which maximize information that can be predicted from knowledge of class membership. This feature is a consequence of COBWEB's design. Contrary to that, in order to select interesting clusters, i.e. formal concepts, the proposed method utilizes an additional information supplied by a user, namely the information regarding characteristic attributes. Therefore, even though both COBWEB and the proposed method result in a system of conceptual clusters, not all but possibly only those which are interesting, the strategies used by the two methods are quite different. A detailed comparison of the proposed method from COBWEB is a topic for future research.

Needless to say, when discussing analysis of questionnaire data, one needs to comment on the possibility to use various methods of statistical data analysis. A general difference of the classic statistical methods from the methods of exploratory data analysis such as formal concept analysis, the method presented in this paper or the methods mentioned above in this section, is that unlike statistical methods, exploratory data analysis tries to retrieve from data some rough information and to suggest possibly interesting patterns and hypotheses from data when there is little known about the data. Contrary to that, statistical methods are used for a more precise data analysis when, for example, a possibly interesting known hypothesis is tested using statistical methods of hypotheses testing. Therefore, statistical methods differ from the methods of exploratory methods in the very aims. Although one could argue that there is no clear distinction between statistical methods and the methods of exploratory nature, the above-mentioned distinction has been recognized a long time ago and is thoroughly discussed e.g. in [23].

In particular, to quantitatively describe the relationship between the characteristics of respondents, methods of regression analysis are appropriate. Such methods have been used in some recent studies, see [12,18,30]. For binary data, logistic regression analysis is appropriate. We ran this method over the 47 attributes described in the Appendix. The strength of association between the level of PA and lifestyle characteristics was examined using SPSS for Windows, version 17.0. The results of logistic regression are presented as odds rations (OR) and 95% confidence intervals (CI). As an example, the likelihood of the model in terms of the Nagelkerke $R2$ value is, for the group of women, $R2 = 0.078$, which corresponds to a rather low (7.8%) ability of the model to explain the data. One of the conclusions drawn from the logistic regression analysis was that a high level of PA of females was significantly ($p < 0.007$) related only to: Walking high (OR = 3.81 [95% CI = 1.80–8.07]) and Job yes (OR = 2.04 [95% CI = 1.22–3.42]). This means that the women with a high level of walking are 3.81 times more likely to have a high level of physical activity compared to the other women. In addition, the women with a job are 2.04 times more likely to have a high level of physical activity compared to the other women. The rest of attributes (living in house, living alone, ownership of car, dog, bike and weekend house, BMI under and normal, younger age) were not significantly related to a high level of PA of females. Similar conclusions were drawn when logistic regression analysis was applied to analyze a group consisting of females and males: a high level of PA was significantly related only to Walking high and Job yes, but again, the ability to explain data measured by the Nagelkerke $R2$ value was rather low ($R2 = 0.097$). A conclusion has therefore been made that the low ability to explain the data does not favor the use of the logistic regression. Namely, in order to be considered a method with a good ability to explain data, the likelihood expressed by means of the $R2$ value is supposed to be at least $R2 = 0.3$.

A further data exploration utilizing classic statistical methods is also one of the topics for future research. As mentioned above, the expert kinathropologist involved in our study was primarily interested in exploratory data analysis. For this purpose, the presented method was considered useful by the expert. Most importantly, the expert appreciated the interactive character and the ability to visualize the data by means of the line diagram of the concept lattice. As a second step, for which various statistical methods would be indispensable, we plan to employ methods of hypotheses testing to be able to draw conclusions saying which of the observations which hold true in the data may be justifiably generalized to whole population.

Acknowledgement

Experiments in this study were carried out using a software for formal concept analysis developed in the Department of Computer Science, Palacky University. The authors thank Vladimír Sklenář, the author of this software, for his help. The authors thank Peter Eklund and the anonymous reviewers for their remarks which helped improve this paper.

Table 5
Guidelines for scaling of data related to physical activity.

	Level of attribute	Overall PA (min. per week)	Walking (min. per week)	Sitting (min. per week)
Aged 15–24 years	Insufficient/low	<199	<59	<3359
	Average	200–399	60–119	3360–5039
	Well–supported/high	>400	>120	>5040
Aged 25–54 years	Insufficient/low	<199	<59	<4199
	Average	200–399	60–119	4200–5879
	Well–supported/high	>400	>120	>5880
Aged 55–65 years	Insufficient/low	<99	<59	<5039
	Average	100–199	60–119	5400–6719
	Well–supported/high	>200	>120	>6720

Appendix A. List of attributes from IPAQ

This section presents a description of the 47 binary attributes derived from the IPAQ questionnaire. The attributes are grouped according to the original IPAQ questions. In addition, Table 5 shows how the actual values of IPAQ data related to physical activity were transformed to the binary attributes.

age—indicates the age of the respondent. This variable was transformed into five attributes:

AGE15–24 – age 15–24 years,
AGE25–34 – age 25–34 years,
AGE35–44 – age 35–44 years,
AGE45–54 – age 45–54 years,
AGE55–65 – age 55–65 years.

overall physical activity—minutes per week for the overall physical activity of the respondent. This variable was transformed into three attributes:

PAlow – insufficient,
PAmoderate – average,
PAhigh – well-supported.

leisure–time physical activity—minutes per week for leisure–time physical activity of the respondent. This variable was transformed into three attributes:

REClow – insufficient,
RECmoderate – average,
REChigh – well–supported.

walking—minutes per week spent on walking by the respondent. This variable was transformed into three attributes according to Table 5:

WALKINGlow – insufficient,
WALKINGmoderate – average,
WALKINGhigh – high.

sitting—minutes per week spent on sitting by the respondent. This variable was transformed into three attributes according to Table 5:

SITTINGlow – low,
SITTINGmoderate – average,
SITTINGhigh – high.

gender—gender of the respondent. This variable was transformed into two attributes:

GENDERmale – man (value 1 in IPAQ),
GENDERfemale – woman (value 5 in IPAQ).

education—level of education of the respondent. This variable was transformed into three attributes:

EDUelementary – basic (0–9 years of education),
EDUsecondary – secondary (10–14 years of education),
EDUhigh – university (at least 15 years of education).

Please cite this article in press as: R. Belohlavek et al., Evaluation of IPAQ questionnaires supported by formal concept analysis, Inform. Sci. (2010), doi:10.1016/j.ins.2010.04.011

job—job status of the respondent. This variable was transformed into two attributes:

JOByes – employed (value 1 in IPAQ),
JOBno – unemployed (value 5 in IPAQ).

height and weight—body height and body weight of the respondent. These data were used to calculate Body Mass Index (BMI) according to the formula: BMI = $body\ weight[kg]/(body\ height[m])^2$. This variable was transformed into four attributes according to the BMI:

BMIunder – underweight (less than 20 kg/m^2),
BMInormal – normal level of body weight (20–24.9 kg/m^2),
BMIover – overweight (25–29.9 kg/m^2),
BMIobesity – obesity (more than 30 kg/m^2).

type of housing—way of living of the respondent. This variable was transformed into two attributes:

HOUSEyes – house (value 1 in IPAQ),
HOUSEno – apartment (value 2 in IPAQ).

smoking—smoking status of the respondent. This variable was transformed into two attributes:

SMOKINGyes – smoker (value 1 in IPAQ),
SMOKINGno – nonsmoker (value 2 in IPAQ).

life—family condition of the respondent. This variable was transformed into three attributes:

LIVINGalone – lives without partner (value 1 in IPAQ),
LIVINGfamily – lives in family without children (value 2 in IPAQ),
LIVINGchild – lives in family with at least one child (value 3 in IPAQ).

material condition—ownership of monitored assets by the respondent. Six attributes were created:

BICYCLEyes – owns bicycle (value 1xx in IPAQ),
BICYCLEno – without bicycle (value 2xx in IPAQ),
CARyes – owns bicycle (value x1x in IPAQ),
CARno – without bicycle (value x2x in IPAQ),
WEEKEND–HOUSEyes – owns weekend house (value xx1 in IPAQ),
WEEKEND–HOUSEno – without weekend house (value xx2 in IPAQ).

organized PA—frequency (times per week) of organized physical activity. This variable was transformed into four attributes:

ORGPAno – without organized PA – value 0,
ORGPAFR1 – one lesson of organized PA per week (value 1 in IPAQ),
ORGPAFR2 – two lessons of organized PA per week (value 2 in IPAQ),
ORGPAFR3+ – more than two lessons of organized PA per week (value 3 in IPAQ).

sporting activity—participation in sport club or sport organization. This variable was transformed into two attributes:

SAno – no participation (value 0 in IPAQ),
SAyes – participation (value 1 in IPAQ).

References

[1] A. Bauman et al, The International Prevalence Study on Physical Activity: results from 20 countries, International Journal of Behavioral Nutrition and Physical Activity 6 (2009) 21, doi:10.1186/1479-5868-6-21.
[2] R. Belohlavek, Fuzzy Relational Systems: Foundations and Principles, Kluwer, Academic/Plenum Publishers, New York, 2002.
[3] R. Belohlavek, Concept lattices and order in fuzzy logic, Annals of Pure and Applied Logic 128 (1–3) (2004) 277–298.
[4] R. Belohlavek, V. Sklenář, J. Zacpal, Crisply generated fuzzy concepts, in: Proceedings of the ICFCA 2005, Lecture Notes in Artificial Intelligence, vol. 3403, Springer-Verlag, Berlin/Heidelberg, 2005, pp. 268–283.
[5] R. Belohlavek, V. Vychodil, Reducing the size of fuzzy concept lattices by hedges, in: Proceedings of the FUZZ-IEEE 2005, The IEEE International Conference on Fuzzy Systems, May 22–25, Reno, Nevada, USA, IEEE Press, 2005, pp. 663–668.

[6] R. Belohlavek, V. Vychodil, What is a fuzzy concept lattice?, in: Proceedings of the CLA 2005, 3rd International Conference on Concept Lattices and Their Applications, September 7–9, Olomouc, Czech Republic, Palacky University Press, 2005, pp 34–45.
[7] R. Belohlavek, V. Vychodil, Attribute implications in a fuzzy setting, in: Proceedings of the ICFCA 2006, Lecture Notes in Artificial Intelligence, vol. 3874, 2006, pp. 45–60.
[8] R. Belohlavek, A note on variable threshold concept lattices: threshold-based operators are reducible to classical concept-forming operators, Information Sciences 177 (15) (2007) 3186–3191.
[9] S. Ben Yahia, A. Jaoua, Discovering knowledge from fuzzy concept lattice, in: A. Kandel, M. Last, H. Bunke (Eds.), Data Mining and Computational Intelligence, Physica-Verlag, 2001, pp. 167–190.
[10] J.-M. Bernard, S. Poitrenaud, Multivariate Bayesian implicative analysis for a binary questionnaire: quasi-implications and simplified Galois lattice (in French), Mathématiques et Sciences Humaines (Mathematics and Social Sciences) 147 (1999) 25–46.
[11] C. Carpineto, G. Romano, Concept Data Analysis. Theory and Applications, J. Wiley, 2004.
[12] E. Cerin, E. Leslie, How socio-economic status contributes to participation in leisure-time physical activity, Social Science and Medicine 66 (2008) 2596–2609.
[13] G.A. Colditz, Economic costs of obesity and inactivity, Medicine and Science in Sports and Exercise 31 (1999) S663–S667.
[14] C.L. Craig et al, International physical activity questionnaire: 12-country reliability and validity, Medicine and Science in Sports and Exercise 35 (2003) 1381–1395.
[15] H. Cutt et al, Dog ownership, health and physical activity: a critical review of the literature, Health and Place 13 (2007) 261–272.
[16] A.-L. Doyen, V. Duquenne, S. Nuques, M. Carlier, What can be learned from a lattice analysis of a laterality questionnaire?, Behavior Genetics 31 (2001) 193–207
[17] D.H. Fisher, Knowledge acquisition via incremental conceptual clustering, Machine Learning 2 (1987) 139–172.
[18] A. Forsyth, J.M. Oakes, B. Lee, K.H. Schmitz, The built environment, walking, and physical activity: is the environment more important to some people than others?, Transportation Research, Part D 14 (2009) 42–49
[19] K. Frömel et al., Physical Activity and Inactivity of Inhabitants of the Czech Republic in the Context of Behavioral Changes (in Czech), RP: 6198959221, Palacký University, Olomouc, Czech Republic, 2005.
[20] K. Frömel, J. Mitáš, J. Kerr, The association between active lifestyle, the size of a community and SES of the adult population in the Czech Republic, Health & Place 15 (2009) 447–454, doi:10.1016/j.healthplace.2008.08.003.
[21] B. Ganter, R. Wille, Formal Concept Analysis. Mathematical Foundations, Springer, Berlin, 1999.
[22] P. Hájek, Metamathematics of Fuzzy Logic, Springer, 1998.
[23] P. Hájek, T. Havránek, Mechanizing Hypothesis Formation. Mathematical Foundations for a General Theory, Springer, 1978.
[24] I. Janssen et al, Comparison of overweight and obesity prevalence in school-aged youth from 34 countries and their relationships with physical activity and dietary patterns, Obesity Reviews 6 (2005) 123–132.
[25] G.J. Klir, B. Yuan, Fuzzy Sets and Fuzzy Logic. Theory and Applications, Prentice-Hall, 1995.
[26] S. Krajči, Cluster based efficient generation of fuzzy concepts, Neural Network World 13 (5) (2003) 521–530.
[27] S. Krajči, A generalized concept lattice, Logic Journal of IGPL 13 (2005) 543–550.
[28] S. Kuznetsov, S. Obiedkov, Comparing performance of algorithms for generating concept lattices, Journal of Experimental and Theoretical Artificial Intelligence 14 (2–3) (2002) 189–216.
[29] R.S. Michalski, R.E. Stepp, Learning from observation: conceptual clustering, in: R.S. Michalski, J.G. Carbonell, T.M. Mitchell (Eds.), Machine Learning: An Artificial Intelligence Approach, Morgan Kaufmann, San Mateo, CA, 1983, pp. 331–363.
[30] J.R. Panter, A.P. Jones, Associations between physical activity, perceptions of the neighbourhood environment and access to facilities in an English city, Social Science and Medicine 67 (2008) 1917–1923.
[31] S. Pollandt, Fuzzy Begriffe, Springer-Verlag, Berlin/Heidelberg, 1997.
[32] J.F. Sallis et al, Neighborhood built environment and income: examining multiple health outcomes, Social Science and Medicine 68 (2009) 1285–1293.
[33] J.F. Sallis et al, Neighborhood environments and physical activity among adults in 11 countries, American Journal of Preventive Medicine 36 (2009) 484–490.
[34] E. Sigmund, J. Zacpal, V. Sklenář, K. Frömel, New approach to evaluation of IPAQ questionnaire physical activity data: formal concept analysis theory, in: 4th International Conference Movement and Health Czech Republic, Palacky University Press, 2005, p. 173, ISBN 80-244-1166-0.
[35] V. Sklenář, J. Zacpal, E. Sigmund, Evaluation of IPAQ questionnaire by FCA, in: Proceedings of the CLA 2005, 3rd International Conference on Concept Lattices and Their Applications, September 7–9, Olomouc, Czech Republic, Palacky University Press, 2005, pp. 60–69.
[36] A. Suchomel, D. Sigmundová, K. Frömel, The role of physical activity in the lifestyle of inhabitants of the Liberec region, Human Movement 9 (2008) 19–26.
[37] J.R. Thomas, J.K. Nelson, S.J. Silverman, Research Methods in Physical Activity, Human Kinetic, Champaign, 2005.
[38] R. Wille, Restructuring lattice theory: an approach based on hierarchies of concepts, in: I. Rival (Ed.), Ordered Sets, Reidel, Dordrecht-Boston, 1982, pp. 445–470.
[39] W.-X. Zhang, J.-M. Ma, S.-Q. Fan, Variable threshold concept lattices, Information Sciences 177 (22) (2007) 4883–4892.
[40] IPAQ, <http://www.ipaq.ki.se/>.