Evaluation of IPAQ questionnaire by FCA

Vladimír Sklenář, Jiří Zacpal, Erik Sigmund

Dept. Computer Science, Palacký University, Tomkova 40, CZ-779 00 Olomouc, Czech Republic
{vladimir.sklenar,jiri.zacpal,erik.sigmund}@upol.cz

Abstract. This paper presents using of Formal Concept Analysis (FCA) in evaluation of IPAQ questionnaire. IPAQ is global epidemiological questionnaire physical activity data. It tries to catch state of physical activity (inactivity) in representative file of population. The goal of authors was find dependencies between demographic data (age, gender, education, occupation, ...) and degree of physical activity. We tried to obtain these dependencies from intents of concept lattice created on the base of questionnaire. Because the whole concept lattice was very large and contained number of concepts not interesting for expert that evaluated data from questionnaire, we used binary relations to constrain it. Primarily, we focused on equivalence relations.

Keywords: FCA, evaluation of questionnaire, constrained concept lattice, equivalence relation

1 Preliminaries and Problem Setting

Evaluation of questionnaire is traditionally way how to discover properties (attributes) shared by important set of respondents (objects) and dependencies between properties of respondents. Standard technique of their evaluation is using statistical methods. In this paper we show another method how to get information from data gained from large set of respondents. We used Formal Concept Analysis (FCA) to evaluate data recorded by more than 4000 respondents in IPAQ questionnaire.

Formal concept analysis In its basic setting, formal concept analysis deals with input data in the form of a table with rows corresponding to objects and columns corresponding to attributes which describes a relationship between the objects and attributes. The data table is formally represented by a so-called formal context which is a triplet $\langle X, Y, I\rangle$ where I is a binary relation between X and Y, $\langle x, y\rangle \in I$ meaning that the object x has the attribute y. For each $A \subseteq X$ denote by $A^{\uparrow}$ a subset of Y defined by

$$A^{\uparrow} = \{y \mid \text{for each } x \in A:\ \langle x, y\rangle \in I\}.$$

Similarly, for $B \subseteq Y$ denote by $B^{\downarrow}$ a subset of X defined by

$$B^{\downarrow} = \{x \mid \text{for each } y \in B:\ \langle x, y\rangle \in I\}.$$

That is, $A^{\uparrow}$ is the set of all attributes from Y shared by all objects from A (and similarly for $B^{\downarrow}$). A formal concept in $\langle X, Y, I\rangle$ is a pair $\langle A, B\rangle$ of $A \subseteq X$ and $B \subseteq Y$ satisfying $A^{\uparrow} = B$ and $B^{\downarrow} = A$. That is, a formal concept consists of a set A (extent) of objects which fall under the concept and a set B (intent) of attributes which fall under the concept such that A is the set of all objects sharing all attributes from B and, conversely, B is the collection of all attributes from Y shared by all objects from A. The set $\mathcal{B}(X, Y, I) = \{\langle A, B\rangle \mid A^{\uparrow} = B, B^{\downarrow} = A\}$ of all formal concepts in $\langle X, Y, I\rangle$ can be naturally equipped with a partial order defined by

$$\langle A_1, B_1\rangle \leq \langle A_2, B_2\rangle \text{ iff } A_1 \subseteq A_2 \text{ (or, equivalently, } B_2 \subseteq B_1).$$

Under $\leq$, $\mathcal{B}(X, Y, I)$ happens to be a complete lattice, called a concept lattice.

We refer to [11] for background information in formal concept analysis (FCA).

Formal concept analysis thus treats both the individual objects and the individual attributes as distinct entities for which there is no further information available except for the relationship I saying which objects have which attributes.

However, in case of evaluation of questionnaire, it is necessary to work with some additional information. First, identity of one concrete object is not interesting (respondents are often anonymous). We want to find out properties common to some subsets of respondents (for example young females). Thus we have to define these interesting subsets and consider only concepts which extent contain all (or majority of) respondents from these subsets. Second, we have to calculate with some noise in data. For example, that small number of respondents have different properties than others in their subsets.

2 IPAQ questionnaire

In 1996, Dr. Michael Booth of Sydney, Australia, initiated a collaborative effort to develop a valid and reliable questionnaire measuring health-related physical activity suitable for both research and surveillance. An international group of physical activity assessment experts were invited to form a working group, referred to as the International Consensus Group for the Development of an International Physical Activity Questionnaire. A year later, the consensus group came together for a meeting at the World Health Organisation (WHO) in Geneva, Switzerland. The purpose of the International Physical Activity Questionnaires (IPAQ) is to provide a set of well-developed instruments that can be used internationally to obtain comparable estimates of physical activity. In response to the global demand for comparable and valid measures of physical activity within and between countries, IPAQ was developed for surveillance activities and to guide policy development related to health-enhancing physical activity across various life domains. In IPAQ questionnaire is many attributes, such as age, gender, education, occupation and other particularities of physical activity (PA) and physical inactivity (PI) at representative file of Czech population between 18 and 65 years old. In 2004 were got data for analysis PA and PI patterns from 2300 women a 2018 men. In respect of much adventitious information

characteristic PA and PI is evaluation by "classical" statistics with hypothesis test almost inexhaustible.

3 Concept lattices of contexts with binary relations

In our recent papers we presented how further information additionally supplied with the basic object-attribute data table can be utilized [2],[5],[6],[7]. We now recall the basic concepts of [6].

Definition 1. *A formal context with a binary relation (R-context, for short) is a structure* $\langle X, Y, I, \equiv \rangle$ *(written also* $\langle \langle X, \equiv \rangle, Y, I \rangle$*) where* $\langle X, Y, I \rangle$ *is a formal context and* $\equiv$ *is a binary relation on* X*.*

Remark 1. (1) We are primarily interested in case when $\equiv$ is an equivalence relation. Then $x_1 \equiv x_2$ means that objects x_1 and x_2 are equivalent from some point of view (similar, indistuinguishable).

(2) Equivalence $\equiv$ may be supplied by an expert or may result from some previous analysis or external source. For example, objects from X may be partitioned by some clustering (based on attributes from Y or some other data available) or some convention (a catalogue). Such a partition gives naturally a rise to an equivalence relation.

If $\equiv$ represents an indistinguishability (or intended indistinguishability), it might be desirable to consider only those formal concepts which do not separate indistinguishable objects. We call such formal concepts compatible.

Definition 2. *For an R-context* $\langle \langle X, \equiv \rangle, Y, I \rangle$*, a formal concept* $\langle A, B \rangle \in \mathcal{B}(X, Y, I)$ *is called compatible with* $\equiv$ *if for each* $x_1, x_2 \in X$*, if* $x_1 \in A$*, and* $x_1 \equiv x_2$ *or* $x_2 \equiv x_1$*, then* $x_2 \in A$*.*

Compatible concepts are thus certain formal concepts from $\mathcal{B}(X, Y, I)$ satisfying a natural restriction with respect to $\equiv$. The set of all formal concepts from $\mathcal{B}(X, Y, I)$ which are compatible with $\equiv$ will be denoted by $\mathcal{B}(\langle X, \equiv \rangle, Y, I)$.

For an equivalence $\equiv$ on X, extents of compatible formal concepts are unions of $\equiv$-classes (recall that an $\equiv$-class corresponding to $x \in X$ is a set $[x]_{\equiv} = \{x' \in X \mid x \equiv x'\}$; the collection of all $\equiv$-classes is denoted by $X/\equiv$).

Theorem 1. *([6])* $\mathcal{B}(\langle X, \equiv \rangle, Y, I)$ *equipped with* $\leq$ *is a complete lattice in which arbitrary infima coincide with infima in* $\mathcal{B}(X, Y, I)$*, i.e. it is a complete* $\bigwedge$*-sublattice of* $\mathcal{B}(X, Y, I)$*.*

It can be shown by an easy example that suprema in $\mathcal{B}(\langle X, \equiv \rangle, Y, I)$ do not generally coincide with suprema in $\mathcal{B}(X, Y, I)$.

4 P%-compatible concepts

When we evaluate questionnaire, we want to find properties that are shared by majority of respondents (or interesting subset of respondents, for example all young females). Thus our previous definition of compatible concept is too strict, because it is unnecessary to desire that attributes in compatible intents are shared by all equivalent objects. In most cases it is sufficient that extent contains only important portion(given in percents) of the class of equivalent objects $[x]_{\equiv}$.

Definition 3. *For an R-context* $\langle\langle X, \equiv\rangle, Y, I\rangle$ *and* $0 \leq p \leq 100$*, a formal concept* $\langle A, B\rangle \in \mathcal{B}(X, Y, I)$ *is called p%-compatible with* $\equiv$ *if for each* $x \in A$*,* $|[x]_{\equiv} \cap A| \geq |[x]_{\equiv}|.p/100$

This is, if object x belongs to extent, than also at least $p\%$ of others objects from the same equivalent class must belong to extent. The set of all formal concepts from $\mathcal{B}(X, Y, I)$ which are $p\%$-compatible with $\equiv$ will be denoted by $\mathcal{B}_p(\langle X, \equiv\rangle, Y, I)$

The following lemma is obvious. It shows a natural result saying that the less percents of objects from $[x]_{\equiv}$ is sufficient, the more formal concepts satisfying the restrictions.

Lemma 1. *If* $p_1 \leq p_2$ *then* $\mathcal{B}_{p_2}(\langle X, \equiv\rangle, Y, I) \subseteq \mathcal{B}_{p_1}(\langle X, \equiv\rangle, Y, I)$

5 Evaluation IPAQ questionnaire by FCA

Creation of context. First step in analyse IPAQ questionnaire by FCA is creation of context from questionnaire data table. The set of objects is set of respondent. The set of attributes is given by queries in questionnaire (age, sex, location, BMI, ...). Because of data are not in bivalent form, we have to transform this date to bivalent form by scaling. The expert provided this data for scaling, who assigned borders between degrees of attribute. For example characteristic age divided to three attributes young (age is less then 20 years), middle (age is between 21 and 55 years) and old (age is more then 55 years). The transformation to context is very important, because bad alignment of borders can make for deformation whole concept lattice. Part of questionnaire is in Fig. 1. Part of context is in Fig. 2..

Resulting context has 72 attributes. We can calculate concept lattice for this context. We have lattice, which has about 21 millions concepts. It is very much for finding dependencies between attributes. Because of it we try to constrain lattice by equivalence relation and consider only p%-compatible concepts.

6 Obtaining equivalence relations

The key question is, how to obtain particular equivalences. The most important is expert, who has to specify, which set of attributes is interesting for him. One

	SEX	AGE	EDUCATION YEARS	EMPLOYED	JOB HOURS	LOCATION	HEIGHT	WEIGHT
1	5	18	9	5	999	2	165	60
2	5	18	12	5	999	4	170	57
3	1	18	11	5	999	3	185	66
4	5	18	12	5	999	3	168	53
5	5	18	12	5	999	3	168	55
6	1	18	3	5	999	3	182	79
7	5	18	12	5	999	3	165	55
8	1	18	12	1	60	2	185	85
9	5	18	11	5	999	4	178	60
10	5	18	9	5	999	2	167	48

Fig. 1. Part of questionnaire

	SEX - Man	SEX - Woman	AGE - young	AGE - middle	AGE - old	EDUCATION - elementary	EDUCATION - secondary	EDUCATION - university	EMPLOYED	UNEMPLOYED	EMPLOYD - not answer	JOB HOURS - few	JOB HOURS - ordinary	JOB HOURS - immoderate	JOB HOURS - not answer	LOCATION - big town	LOCATION - medium-sized town	LOCATION - small town	LOCATION - village	LOCATION - not answer	BMI - low	BMI - normal	BMI - high	BMI - obesity	BMI - not answer
1	0	1	1	0	0	1	0	0	0	1	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0
2	0	1	1	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0
3	1	0	1	0	0	0	1	0	0	1	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0
4	0	1	1	0	0	0	1	0	0	1	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0
5	0	1	1	0	0	0	1	0	0	1	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0
6	1	0	1	0	0	1	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0
7	0	1	1	0	0	0	1	0	0	1	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0
8	1	0	1	0	0	0	1	0	1	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0
9	0	1	1	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0
10	0	1	1	0	0	1	0	0	0	1	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0

Fig. 2. Part of context

class of equivalent objects then contains all objects (respondents) that have the same subset of interesting attributes.

More formally. For a formal context $\langle X, Y, I \rangle$ and set of interesting(important) attributtes $M \subset Y$ we denote by $\equiv_M$ the binary relation defined on X by

$x_1 \equiv_M x_2$ if and only if $\{x_1\}^{\uparrow} \cap M = \{x_2\}^{\uparrow} \cap M$.

In other words, $x_1 \equiv_M x_2$ if and only if x_1 and x_2 have the same subset of attributes from M. Obviously, $\equiv_M$ is an equivalence relation on X.

In our case, expert was interested in discavering attributes, that are common for all (or important part of) respondents from given class (for example smoking old men).

Together with expert we defined 32 sets of important attributes, which are from 4 main groups:

- Physical activity, age and gender of respondents.
- Physical activity, age, gender and education of respondents.
- Physical activity, age, gender and body mass index (BMI) of respondents.
- Physical activity, age, gender smoking of respondents.

All above mentioned attributes are many-valued attributes with nominal scale. Each set of many-valued attributes built up equivalent classes of respondents, which have value of this attributes identical. We calculated constrained concept lattices for each equivalence relation. Because of the data from respondents are very sensitive for noise, we also built up lattices contained 90% - compatible and 75% - compatible concepts. We delivered these constrained lattices to the expert to analyze. He finds "interesting" concepts, which describe dependencies between demographic data and physical activity or inactivity. Each compatible concept is interesting for his intent, which contains at most one value for each many-valued important attribute. In addition to these attributes may be in intent contained another attributes. Occurrence of such attributes is interesting for expert, because they are shared by majority of respondents from given equivalence class. Cardinality of extent is also interesting, because it determines number of respondents, who have attributes in intent.

We demonstrate this method on one group of attributes.

7 Example

Expert selected those attributes: gender, age, education and intensive physical activity (PA). Because of gender is scaling to 2 attributes(Man, Woman), age to 3 attributes(young, middle, old) and intensive PA to 3 attributes(below-average, average, above-average) we have 54 equivalent classes. Corresponding constrained concept lattices have 188 concepts. Set of all p%-compatible concepts contain 418 concepts (90%) and 1 449 concepts (75%). Now the expert can analyze lattices. He choose one equivalent class. For example: SEX - man,

AGE - middle, EDUCATION - secondary, intensive physical activity (PA) - above-average. For those attributes he finds greatest(by concept lattice ordering) concept, which intent contains all this attributes. Such concept with all concepts, which are less create sublattice. Expert goes through this sublattice and finds out intents, which contains another attributes then those, which characterize given class(or group of classes). At first we analyze sublattice, which contains 100%-compatible concepts. Corresponding sublattice is in Fig. 3.

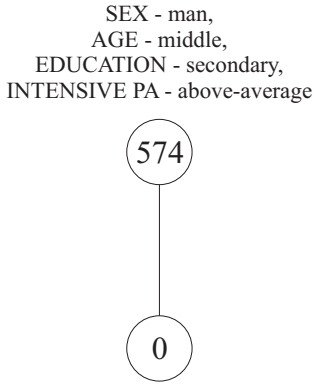

Fig. 3. sublattice contained 100% - compatible concepts

This sublattice has only smallest and greatest element. There is not another concept in this sublattice. It is causing by requirment, that all objects-respondents from equivalent class have to be contained in concept. For expert is important intent of greatest concept, because in it are all atribututes common for all respondents from given class. In this example we can see, that there are only attributtes, which determine given equivalente class. More interesting is lattice contained 90% - compatible concepts. Corresponding sublattice is in Fig. 4.

There are 3 aditionally concepts. First includes 569 respondents (it is 99% from all members of equivalent class), who have value of attribute SITTING equal to low. It confirm expecting, that respondents with high intensive PA sitting low. Second concept includes respondents, who have value of attribute NATIONALITY equal to Czech. This fact we would interprete that only Czech respondents have high intensive PA. Really it means, that majority of all respondents had Czech nationality. Third concept is infimum of previous two concepts.

The largest sublattice is from lattice contained 75% - compatible concepts. Corresponding sublattice is in Fig. 5.

This sublattice has some interesting concepts. For example the concept, which include 457 respondents (it is 79% from all members of equivalent class) with value of attribute SPORT ACTIVITY equal to yes, value of attribute SITTING equal to low and value of attribute NATIONALITY equal to Czech. We can de-

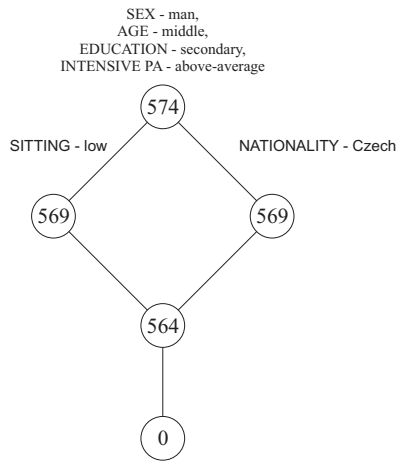

Fig. 4. sublattice contained 90% - compatible concepts

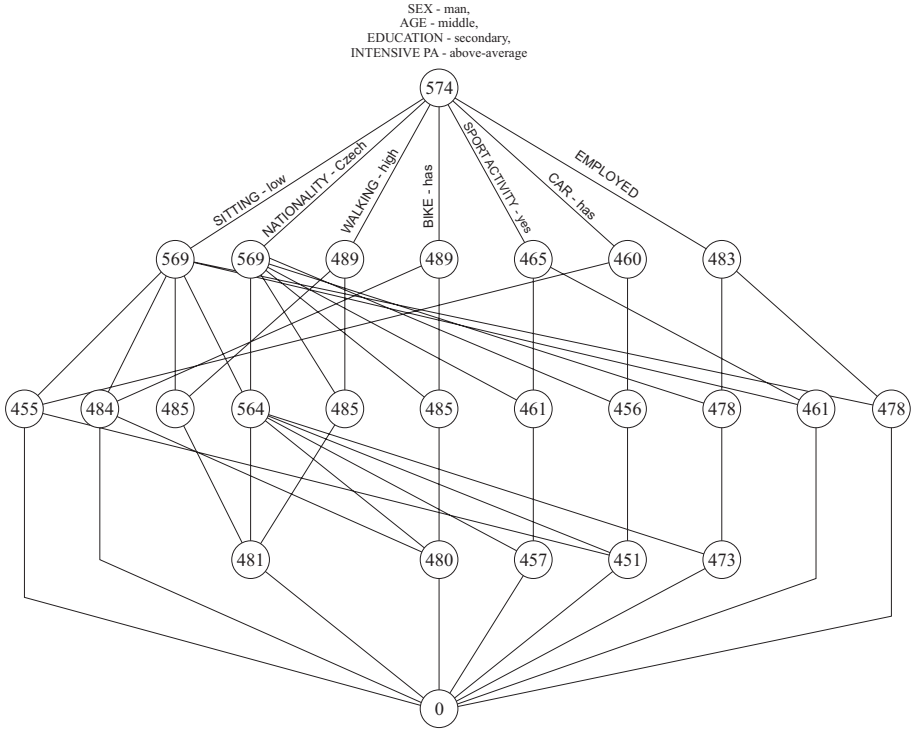

Fig. 5. sublattice contained 75% - compatible concepts

duce form this concept, that above-average physical activity is closely associated with fact, that person sits low and does some sport activity during his leisure time.

8 Future research

We now comment on some further topics and future research (some of these are studied in [6]).

- A concept lattice may be thought of as a hierarchical clustering scheme. The partition corresponding to $\equiv$ represents another clustering (more generally, we can think of a hierarchical clustering scheme). Several interesting problems arise here (constraining one clustering by the other, comparing the clusterings, measuring their mutual consistency, etc.), a work is in progress.
- There is more ways of creating context from questionnaire. Naturally way is using of fuzzy logic and fuzzy sets. We will experiment with creating of fuzzy context and methods of constraining resulting fuzzy concept lattice by fuzzy relations. Main ideas of fuzzy concept analysis are in [3],[4].

9 Conclusion

Our way of evaluating gives a new point of view on data contained in questionnaire. On the base of first response from expert, who worked with our results, we can say, that our approache may be usefull for finding dependencies between properties of respondents of questionnaire.

Acknowledgment Supported by grant No. 1ET101370417 of the GA AV CR.

References

1. G. Ammons, D. Mandelin, R. Bodik, J. R. Larus. Debugging temporal specifications with concept analysis. In *Proc. ACM SIGPLAN'03 Conference on Programming Language Design and Implementation,* pages 182–195, San Diego, CA, June 2003.
2. R Bělohlávek, V. Sklenář, J. Zacpal. Formal concept analysis with hierarchically ordered attributes. *Int. J. General Systems* 33(4)(2004), 283-294.
3. Bělohlávek R.: *Fuzzy Relational Systems: Foundations and Principles.* Kluwer Academic/Plenum Publishers, New York, 2002.
4. Bělohlávek R.: Concept lattices and order in fuzzy logic. Annals of Pure and Applied Logic 128(1-3)(2004), 277-298.
5. Bělohlávek R., Sklenář V.: Formal Concept Analysis Constrained by ADF. In: Proc. ICFCA 2005, pp. 176–191. [ISBN 3-540-24525-1]
6. Bělohlávek R., Sklenář V., Zacpal J.: Concept lattices constrained by equivalence relations. In: Proc. CLA 2004, pp. 58–66. [ISBN 80-248-0597-9]
7. Bělohlávek R., Sklenář V., Zacpal J.: Concept lattices constrained by systems of partitions. In: Proc. Znalosti 2005, pp. 5–8.

8. C. Carpineto, R. Romano. A lattice conceptual clustering system and its application to browsing retrieval. *Machine Learning* 24:95–122, 1996.
9. R. Cole, P. Eklund. Scalability in formal context analysis: a case study using medical texts. *Computational Intelligence* 15:11–27, 1999.
10. U. Dekel, Y. Gill. Visualizing class interfaces with formal concept analysis. In *OOPSLA'03*, pages 288–289, Anaheim, CA, October 2003.
11. B. Ganter, R. Wille. *Formal Concept Analysis. Mathematical Foundations.* Springer-Verlag, Berlin, 1999.
12. O. S. Kuznetsov, S. A. Obiedkov. Comparing performance of algorithms for generating concept lattices. *J. Exp. Theor. Artif. Intelligence* 14(2/3):189–216, 2002.
13. D. Maier. *The Theory of Relational Databases.* Computer Science Press, Rockville, 1983.
14. O. Ore. Galois connections. *Trans. Amer. Math. Soc.* 55:493–513, 1944.
15. G. Stumme, R. Wille, U. Wille. Conceptual knowledge discovery in databases using formal concept analysis methods. In J. M. Zytkow, M. Quafofou (Eds.). *Principles of Data Mining and Knowledge Discovery.* LNAI 1510, pages 450–458, Springer, Heidelberg, 1998.
16. P. Valtchev, R. Missaoui, R. Godin, M. Meridji. Generating frequent itemsets incrementally: two novel approaches based on Galois lattice theory. *J. Exp. Theor. Artif. Intelligence* 14(2/3):115–142, 2002.