Communication papers of the Federated Conference on Computer Science and Information Systems, pp. 9–16 DOI: 10.15439/2017F216 ISSN 2300-5963 ACSIS, Vol. 13

Data Mining with Trusted Knowledge

Viktor Nekvapil
University of Economics, Prague
nam. W. Churchilla 4
130 67 Praha 3
Email: viktor.nekvapil@vse.cz

□

Abstract—**In this paper, a new concept of Trusted Knowledge (TK) is introduced. Trusted Knowledge are data from trusted organizations such as ministries, statistical offices and so on which can replace a domain expert in the evaluation phase of the data mining task. Two approaches to applying Trusted Knowledge are introduced. The first one called "Explanation system" offers additional information relevant to the resulting patterns which can help the user to better understand results of the task. The second one called "A/TK-formulas" filters out the resulting patterns which are consequences of Trusted Knowledge and thus enables the user to concentrate on the interesting patterns. Conversely, the user can request to be shown only the resulting patterns which are consequences of TK to see which of them are in line with TK. Feasibility of the newly proposed framework is demonstrated in a case study.**

I. Introduction and Related Work

As STATED in the paper "10 challenges of data mining research" [1], "there is a strong need to integrating data mining and knowledge inference". Although there have been some achievements since the paper has been published (see e.g. [2], [3]), data mining systems are still "unable to relate the results of mining to the real-world decisions they affect", as the authors claimed. Moreover, they stated that "Doing these inferences, and thus automating the whole data mining loop requires representing and using world knowledge within the system. One important application of the integration is to inject domain information and business knowledge into the knowledge discovery process".

The approach presented in my paper contributes to this challenge. It incorporates additional knowledge in the evaluation phase of data mining but avoids a lengthy and complex task of building a belief system of the user (see e.g. [4], [5], more recently in [6]). The idea is to enhance user's domain knowledge using available trusted sources of data – that is, data from trusted organisations such as statistical offices, ministries and so on. I refer to this knowledge as ***trusted knowledge***.

As domain experts are often not at disposal or are costly, the research aim is to replace to some extent the domain expert with automate means – the system which will offer additional information to the user instead of the domain expert. Summarisingly, using knowledge and experience of himself/herself, the user evaluates whether the resulting pattern is interesting or not, or if he/she does not have enough knowledge in the particular domain, he/she uses the knowledge from the domain expert.

Further related work includes the approaches which use linked open data (LOD), which allow for publishing interlinked datasets employing machine interpretable semantics. For example, [7] developed an extension for Rapid Miner, which can extend a dataset with additional attributes drawn from the Linked open data cloud. Many approaches (e.g. [8], [9]) use DBPedia as a source of knowledge when evaluating resulting patterns; however, in my opinion, the nature of the data – it comes from *community* – is not trustworthy source. Reference [10] traverses Linked Data to find commonalities that form explanations for items of a cluster. However, the feature of LOD-based approaches is the fact that one has to map data to the ontology first; on contrary, the approach presented in this paper does not require lengthy setup.

The concept of Trusted Knowledge is inspired by FOFRADAR framework [3]. FOFRADAR is based on a logical calculus of association rules. The interpretation is based on mapping important items of knowledge to the sets of association rules which can be considered as their consequences. Important items of knowledge are expressed using a simple mutual influence among attributes. These are predefined relationships of attributes which are used to determine whether the association rule can be seen as a consequence of the item of knowledge or not. For example, the simple mutual influence (SI-formula) *Income* ↑↑ *Loan* means: "*if Income increases, then Loan increases as well*". The set of atomic consequences of this SI-formula can be expressed by the following union: $Low_{Income} \times Low_{Loan} \cup Medium_{Income} \times Medium_{Loan} \cup High_{Income} \times High_{Loan}$, saying that "*if Income is high, than Loan is high or if Income is medium then Loan is medium or if Income is high then Loan is high*". Based on the *levels* in the union, it is possible to say whether the resulting rule is a consequence of the defined *SI-formula* or not. This feature is used in the

□ The work described here has been supported by the internal grant agency of the University of Economics, Prague under project IGA 29/2016.

proposed framework and further developed, as obvious in the following sections.

Theoretical concepts in this paper are demonstrated on a real data set from a financial institution. There are data concerning clients, who were given a loan, including geographical and demographical client data, data from the loan application, data concerning the agent who arranged the loan, and so on.

The rest of the paper is organised as follows. In section II, Trusted Knowledge is defined. Two possible ways of applying Trusted Knowledge are described in section III. In section IV, a case study is presented. In section V, conclusions and suggestions for the future work are included.

II. Trusted Knowledge

A. *Sources of Trusted Knowledge*

There are various data publicly available that can be used as Trusted Knowledge. Government institutions, EU institutions and statistical offices offer more and more data. This is boosted by the Open government data initiatives (see *http://opengovernmentdata.org/*), which offer a catalogue of publicly available data sets. In the Czech Republic, the Open data initiative (see *http://www.opendata.cz/en*) offers a catalogue of data using the linked data paradigm which refers to the Czech Republic. The data from those organisations are generally considered to be trusted sources.

I define Trusted Knowledge as follows: ***Trusted Knowledge*** (TK) is the data from trusted sources which can be connected to the results of a data mining task and are used in the evaluation phase of the data mining task to help with the understanding of the results. Trusted Knowledge can be seen as a special case of domain knowledge.

Trusted Knowledge is obtained from a trusted organisation. An example of such knowledge is the average and median income per district in the Czech Republic obtained from Czech Statistical Office [11].

B. *Items of Trusted Knowledge*

The following items of Trusted Knowledge are defined – *measures of TK*, *levels of measures of TK*, *explanations* and a *mutual influence of the attribute and measure of TK (A/TK-formula)*. The first two items are discussed in the following sections, the remaining items are described in section III.

I. *Measures of TK*

Measure of Trusted Knowledge (measure of TK) is a formalised piece of Trusted Knowledge. I formalise the measure of TK as follows:

a) Each measure has its name; b) is stored as per another dimension – in our case per geographical dimension; c) each value of dimension has its rank within the measure stored; d) each value of dimension has the absolute value of the measure stored.

An example of the measure of TK is depicted in Table III.

In Fig. 1, I outline the basic feature of the measure of TK – its close connection to the results of a data mining task (resulting patterns). I use association rules as an example. Geographical dimension (locality) is used as a *connecting element* between the measure of TK and resulting patterns.

An average income in District X as a measure of TK and *The loan amount taken by a client in District X* as an attribute from analysed data can be examples of such a connection. If such a connection is done, it is assumed that the client is a member of the population which has an average income amounting to 20456, because he or she lives in the same district as the people whose income was collected by a trusted organisation (Czech Statistical Office in this case).

Of course, there are some challenges for this assumption, as for example, when the client lives in a particular region but works in a different one. Nonetheless, I believe that this situation is not occurring frequently and the given principle can be used in general.

To distinguish between data and *Trusted Knowledge*, I use the term *attribute* for the variables derived from the analysed data and the *measure of TK* for the variables used as *Trusted Knowledge*. Note that both the measure of TK and the attribute connected via a *connecting element* are ordinal.

II. *Levels of measures of TK*

The relationship depicted in Fig. 1 above does not bring much insight on its own. It is necessary to bring more context to the relationship. ***Levels of measures of TK*** enables to easily compare attributes and measures of TK. The way how domain experts evaluate the found patterns is commonly expressed by easily interpretable phrases saying for example *"Income is low"*, *"Amount is high"* and so on. This simple approach is followed in FOFRADAR, as described in section I.

Recall the set of atomic consequences of SI-formula $Income \uparrow\uparrow Loan$: $Low_{Income} \times Low_{Loan} \cup Medium_{Income} \times \times Medium_{Loan} \cup High_{Income} \times High_{Loan}$. Now we have to define what means, for example, *"Income is low"* (that is to define the level Low_{Income}).

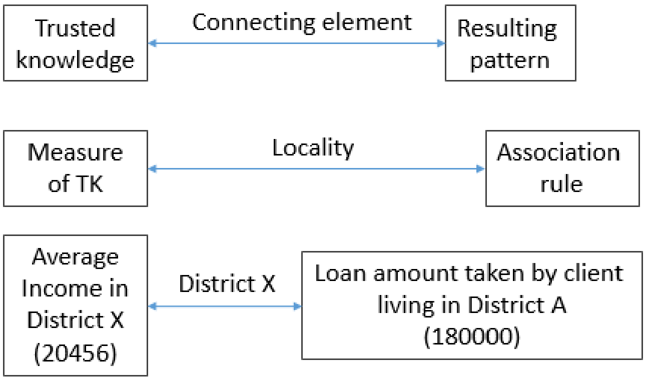

Fig. 1: Relationship between measure of TK and resulting patterns – different degrees of abstraction

The task is to assign a set of values (called *categories* in FOFRADAR) α_l of a particular attribute *A* to each level *Lev(l)*. Levels have scales of various length – for example, if l=3, then levels Lev(l)={*low, medium, high*} or if l=5, then levels Lev(l) = {*very low, low, medium, high, very high*} and so on.

Now, I will present two approaches of defining levels, one of which is the newly proposed *Rank-based approach*.

Expert-based approach means that the domain expert decides which category is assigned to each *level*. For example, he/she can decide that *Level=low* for attribute *Loan* can be considered for the following categories: <0;100000),<100000;150000). Formally written, Low_{Loan} = {<0;100000), <100000;150000)}. This approach is now used in the FOFRADAR framework and its feature is that it is necessary to define sets of categories with the help of the domain expert.

Rank-based approach is the newly proposed way of the automatic definition of *levels*. Categories of a particular attribute or measure of TK are sorted from the lowest to the highest. Then, we assign rank to each of the category according to the value of attribute or measure of TK. Last step comprises assigning *Level(l)* to each rank. For example, consider the categories of the attribute *Loan_amount* depicted in Table I. Based on the rankings of the categories, it is possible to assign respective categories to the levels.

This can be done by applying *Assignment rules* which state the principles of assigning categories to levels. Assignment rules are generic and not dependent on the analysed data. For example, assignment rules can look as depicted in Table II.

Table I: Levels for attribute Loan_amount

Loan_amount category	**Rank**	**Level**
<0; 100000)	1	Very low
<100000; 150000)	2	Very low
<150000 ;200000)	3	Low
<200000; 270000)	4	Low
<270000; 300000)	5	Medium
<300000; 400000)	6	Medium
<400000; 500000)	7	High
<500000; 550000)	8	High
<550000; 650000)	9	Very high
<650000; 2600000>	10	Very high

Table II: Assignment rules

#	**Number of levels α**	**Number of categories**	**Assignment rule**
1	5	10	1 level per 2 categories
2	5	14	1 level per 4 categories, (top and bottom levels per 3 categories), overlapping levels

Table III: Levels for measure of TK Income

District	**Income**	**Income rank**	**Level**
Hlavni mesto Praha	35 115	1	Very high
Stredocesky kraj	27 345	2	Very high
Jihomoravsky kraj	26 116	3	Very high/High
Plzensky kraj	26 026	4	High
Moravskoslezsky kraj	24 877	5	High
Liberecky kraj	24 767	6	High/Medium
Kralovehradecky kraj	24 387	7	Medium
Ustecky kraj	24 336	8	Medium
Jihocesky kraj	24 321	9	Medium/Low
Kraj Vysocina	24 293	10	Low
Olomoucky kraj	24 175	11	Low
Pardubicky kraj	24 067	12	Low/Very low
Zlinsky kraj	23 873	13	Very low
Karlovarsky kraj	22 707	14	Very low

First rule says that 2 consecutive categories are contained in 1 level. This assignment rule is applied in Table I. Second rule is example of overlapping levels, one category could be assigned to two levels. This behaviour is demonstrated in Table III on the measure of TK *Income*.

It is possible to prepare categories of attribute in such a way that it is easy to assign levels to each category. That means, if one desires 5 levels of an attribute, one creates 5 or 10 categories of the attribute, and similarly, this applies when 3 levels are considered, and so on.

Having the levels of attributes and measures of TK defined, we can compare levels and draw consequences based on values of the levels. This is further elaborated upon in section III.

III. Applying trusted knowledge

Here, I will introduce ***Trusted Knowledge Framework*** (TK Framework) – a framework which shows how TK is applied in the data mining process.

An important component of the framework is *Trusted Knowledge Repository* (TKR), a database where the items of TK are stored. Its feature is that it will be possible to share TKR among different projects in a similar way, as for example in [17]. The principles of sharing are left for the future work.

Now, I define two approaches of applying TK. The first approach is less demanding on the domain knowledge that has to be defined in advance but enables less automation in the evaluation phase. I call this approach *Explanation system* and elaborate upon this in section III.A. The second approach follows closely the FOFRADAR principles of automatic conclusions. It is called *A/TK-formulas* and is discussed in section III.B.

A. Explanation system

Although the term *explanation* is broadly used in relation to expert systems (see e.g. [12]), I use it here in connection with a data mining system. Here, I perceive ***explanation*** as an item of TK that could help the user of data mining system to better understand results of a data mining task.

This is especially useful in situations when it is hard to obtain relevant knowledge from domain experts. In this case, a relevant explanation can be used as a support for the user and no knowledge from domain experts is needed.

Explanation is based on the measure of TK. As an example, I will mention the following explanation based on the measure of TK *Income* from Table III:

Zlinsky kraj => Income (very low),

meaning that in *Zlinsky kraj* district, *Income* is *very low*.

The TK Framework specified for the Explanation system is depicted in Fig. 2. As can be seen, after the results are obtained from the data mining system, TKR is queried for relevant explanations. If found, the relevant explanations are handed back to the user. Additionally, the user can request *context* of the explanation to better understand the explanation.

Now I will describe the experimental implementation of the proposed approach. It is a semi-automatic implementation based on LISp-Miner, SQLite and Python. The LISp-Miner System has been chosen for its ability to fine-tune the set of association rules which is mined; for more details see [13]. Further advantage is the possibility to automate the task through LMCL scripting language [14]. A simple database table is used as TKR. SQLite database is used for this purpose. Python is used as an engine that retrieves data from the TKR and presents it in the form of an explanation for the user.

I will continue using the following example. The 4ft-Miner procedure of the LISp-Miner is employed. I define the task so that attribute *District* is in the antecedent of the resulting rules, in consequent, *Loan_amount* is present.

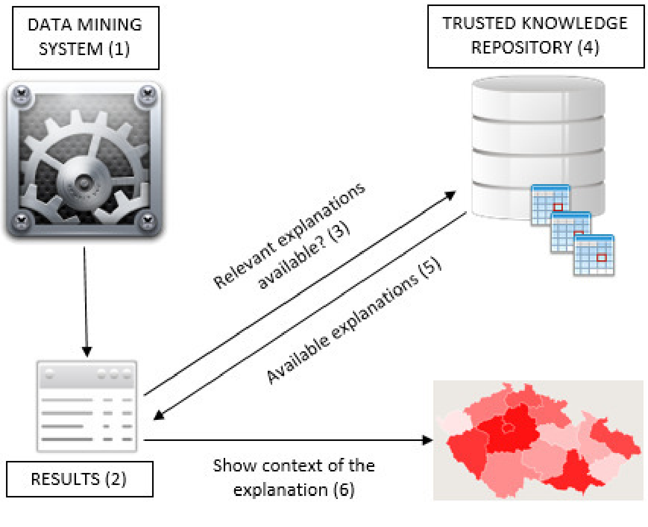
DATA MINING SYSTEM (1)
TRUSTED KNOWLEDGE REPOSITORY (4)
Relevant explanations available? (3)
Available explanations (5)
RESULTS (2)
Show context of the explanation (6)

Fig. 2: TK framework of the Explanation system

In the TK Framework (see Fig. 2), this activity corresponds to number 1 (*Data mining system* and its usage). Three rules were found which are relevant to the task definition; see Fig. 3. This step corresponds to number *2 - Results* in the TK framework.

If we take, for example, rule (1), the user knows from the distribution of the attribute *Loan* (see Table I) that *Loan <100000; 150000)* is rather low (has the level *very low*). Are there some explanations for this fact? This question corresponds to the activity number 3 in the TK Framework. The TKR contains data from trusted organisation – Czech Statistical Office. The average income as presented in Table III above is contained in the TKR. Relevant explanations will be handed back to the user. The explanations which signal that a district presented in the resulting rule is outstanding (unusual) as regards the selected measure of TK, is considered to be relevant. The relevancy criterion in the experimental implementation is defined as the value of geo dimension having a *very high* or *very low* level according to the selected measure of TK. The example output of Explanation system for the first rule is presented below (see Fig. 4). First of all, basic information about the rule is summarised. Then, the explanations found are presented.

The user can conclude from the explanation that the average income in district *'Zlinsky kraj'* is rather low (*very low* level). He/she can conclude that the clients' data he/she has at disposal in the data set are in line with his/her domain knowledge, because there is a direct proportion between the amount of loan and income (which is a part of the users' domain knowledge).

The reasoning used here is a combination of domain knowledge of the user and the measure of TK which brings the user to the conclusion that the rule is in line with his domain knowledge. The user knows that there is a direct proportion between the amount of loan and income, but does not know if the income in the district (*'Zlinsky kraj'*) is rather low or high.

(1) District (Zlinsky kraj) -> Loan_amount <100000; 150000)
(2) District (Hlavni mesto Praha) -> Loan_amount <500000; 550000)
(3) District (Hlavni mesto Praha) -> Loan_amount <550000; 650000)

Fig. 3: Resulting rules

```
---------------------------------------------
Rule ID: 34
Rule: District(Zlinsky kraj) ->
Loan_amount<100000;150000)
Lift: 1.68571
Support: 0.0176
Geo attribute found: District
Coefficient of geo attribute: Zlinsky kraj
Explanations found:
--- Explanation 1 ---
Zlinsky kraj => income_avg (very low)
Value of the geo dimension: Zlinsky kraj
Measure: income_avg
Level of the measure: very low (bottom 2)
Value of the measure: 23873
---------------------------------------------
```

Fig. 4: Example output of the Explanation system

This reasoning can be further automated as presented in section III.B. At this stage, the context of the explanation (as depicted in Fig. 2) can be obtained in a very simple manner. A table with districts, values of the measure, rank and level is retrieved. As a future work, a map with all the data mentioned above will be retrieved to enable to see values of neighbouring districts.

B. *Consequences of Trusted Knowledge (A/TK-formulas)*

One of the possible solutions of the automatic formulation of conclusions using domain knowledge is presented in the FOFRADAR framework, as described above. Using the measures of TK, it is possible to define mutual influence between an attribute and measure of TK. I call this mutual influence ***Attribute / Trusted Knowledge-formula*** **(*A/TK-formula*)** and consider it an *item of Trusted Knowledge*. The principle of A/TK-formula is the same as in FOFRADAR, but instead of one of the attributes, the measure of TK is used in the mutual influence.

The TK Framework of A/TK-formulas is depicted in Fig. 5. After results are obtained, TKR is queried for A/TK-formulas (3) which are available and relevant for the resulting patterns. The formulas are returned to (5) and their consequences are applied to the resulting patterns (6). Alternatively, the user can define a new A/TK-formula in advance using measures of TK that is available in TKR and then, consequences can be applied.

There are two ways how the consequences of A/TK-formulas can be applied:

I. to obtain patterns which are consequences of A/TK-formula – this way is useful when the user wants to know which resulting patterns are in line with the overall knowledge (trusted knowledge)

II. to filter out patterns which are consequences of A/TK-formula – this way the user can filter out resulting patterns which are in line with trusted knowledge and concentrate on patterns which are not consequences of TK (they are either in contradiction to TK or have no TK available).

As an example, let us discuss the A/TK-formula *Income* ↑↑ *Loan*. *Income* is a measure of TK. Using the rank-based approach, it is possible to assign values to respective levels as shown in Table III. The categories of the attribute Loan can be assigned to the levels, as depicted in Table I. Then the set of consequences of the A/TK-formula *Income* ↑↑ *Loan* is defined by the following union:

$$\mathit{Very\ low}_{\mathit{INCOME}} \times \mathit{Very\ low}_{\mathit{LOAN}} \cup \mathit{Low}_{\mathit{INCOME}} \times \mathit{Low}_{\mathit{LOAN}} \cup \mathit{Medium}_{\mathit{INCOME}} \times \mathit{Medium}_{\mathit{LOAN}} \cup \mathit{High}_{\mathit{INCOME}} \times \mathit{High}_{\mathit{LOAN}} \cup \mathit{Very\ high}_{\mathit{INCOME}} \times \mathit{Very\ high}_{\mathit{LOAN}}$$

Using the 4ft-Miner procedure, the following 8 rules are results of the task (see Fig. 6). An important feature of the 4ft-Miner procedure is that it is possible to define sequences of coefficients. In our example, *Loan_amount* was defined as a sequence of max. length 3.

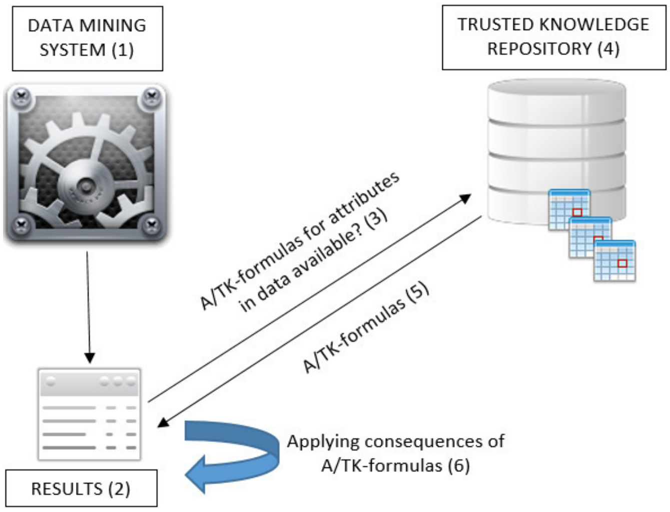
DATA MINING SYSTEM (1)
TRUSTED KNOWLEDGE REPOSITORY (4)
A/TK-formulas for attributes in data available? (3)
A/TK-formulas (5)
Applying consequences of A/TK-formulas (6)
RESULTS (2)

Fig. 5: TK framework of A/TK-formulas

This means that up to 3 consecutive categories can be chained together to increase the support of the rule; see [13] for details on sequences. This behaviour is performed on-the-fly when the 4ft-Miner is running. Here, I do not state the measure of interestingness due to space reasons.

For rule (1), Loan has category *<100000; 150000*), so the assignment rule assigns the level '*very low'* to the attribute. Now, the connecting element District is used to link the rule to the measures of TK., in our case, to the *Income* measure. The district value is *Zlinsky kraj.* If one looks at the level of the measure *Income*, it is *very low* according to the assignment rule in Table III. Note that in some rules, for example (4), the sequence option resulted in broader intervals. For example *Loan_amount (>= 500000*) stands for the following categories: *<500000; 550000*), *<550000; 650000*), *<650000; 2600000>*. It is possible to determine levels of the attribute *District* and the measure of TK *Income* for each rule, as can be seen in Table IV.

Then, we can apply consequences of the A/TK-formula *Income* ↑↑ *Loan*. It is possible to filter out consequences of the formula, that is, rules 1 and 5 and reduce the number of displayed rules from 8 to 6. Conversely, one can request all rules which are consequences of the A/TK-formula *Income* ↑↑ *Loan* and display only rules 1 and 5. The idea is further elaborated upon in section 4.

The novelty of the approach is that the A/TK-formula contains only one attribute of the data analysed, the second element is a measure of TK. Another important feature is that a connecting element is used to link up the measure of TK and the attribute in the resulting patterns.

(1) District (Zlinsky kraj) -> Loan_amount (<100000; 150000)
(2) District (Hlavni mesto Praha) -> Loan_amount (<500000; 550000)
(3) District (Hlavni mesto Praha) -> Loan_amount (<500000; 650000)
(4) District (Hlavni mesto Praha) -> Loan_amount (>= 500000)
(5) District (Zlinsky kraj) -> Loan_amount (<150000)
(6) District (Olomoucky kraj) -> Loan_amount (<270000; 400000)
(7) District (Zlinsky kraj) -> Loan_amount (<100000; 200000)
(8) District (Hlavni mesto Praha) -> Loan_amount (<400000; 550000)

Fig. 6: Resulting association rules

Table IV: Categories and corresponding levels for each rule

#	*Loan* category in rule	*Loan* level	Connecting element value	*Income* level	Cons. of *Income* ↑↑ *Loan*
1	<100000; 150000)	Very low	Zlinsky kraj	Very low	Yes
2	<500000; 550000)	High	Hlavni mesto Praha	Very high	No
3	<500000; 650000)	High AND Very high	Hlavni mesto Praha	Very high	No
4	<500000; 2600000)	High AND Very high	Hlavni mesto Praha	Very high	No
5	<0; 150000)	Very low	Zlinsky kraj	Very low	Yes
6	<270000; 400000)	Medium	Olomoucky kraj	Low	No
7	<100000; 200000)	Very low AND Low	Zlinsky kraj	Very low	No
8	<400000; 550000)	High	Hlavni mesto Praha	Very high	No

IV. Case study

The case study makes use of the same data that were used in the examples in the above sections. The goal of the case study is to show a more complex example of using the proposed framework. The combination of Explanation system and A/TK-formulas is shown as well as the relationship between the definition of categories of rational attributes (generally known as binning) and the assigning of levels to the categories of attributes.

After discussions with business experts, the task was specified as follows:

Are there any interesting combinations of client properties and indicators (including Loan_amount) on one side and locality on the other?

I am interested in filtering out rules which are consequences of A/TK-formulas contained in TKR. I also want to obtain explanations for the rules.

A. Items of Trusted Knowledge in TKR

In TKR, the following measures of TK are contained.

- *Average income per district* (see Table III)
- *Average price of flat per square meter per district* (see Table V) – source [15]
- *Average amount of mortgage per district* (Table VI) – source [16]

The following A/TK formulas are contained in TKR:

- *Loan_amount* ↑↑ *Price of flat* – if the amount of a loan is high, then the price of a flat is also high
- *Loan_amount* ↑↑ *Income* - If the amount of a loan increases, then the income increases as well.

The relevancy criterion for explanations is set to *District – 'very high'* or *'very low'*. The connecting element is the attribute / dimension District.

Table V: Levels for measure of TK Price of flat

District	Price per square meter	Rank	Level
Hlavni mesto Praha	61500	1	Very high
Jihomoravsky kraj	46800	2	Very high
Kralovehradecky kraj	37100	3	Very high/High
…	…	…	…
Karlovarsky kraj	22100	12	Low/Very low
Moravskoslezsky kraj	18400	13	Very low
Ustecky kraj	10700	14	Very low

Table VI: Levels for measure of TK Average amount of mortgage (mortgage_avg)

District	Mortgage amount (mil.)	Rank	Level
Hlavni mesto Praha	2.721	1	Very high
Jihomoravsky kraj	1.933	2	Very high
Plzensky kraj	1.806	3	Very high/High
…	…	…	…
Zlinsky kraj	1.59	12	Low/Very low
Kraj Vysocina	1.542	13	Very low
Karlovarsky kraj	1.467	14	Very low

B. Task definition in LISp-Miner

The *'Clients' properties'* group of attributes include the attributes *Bonity*, *Collection*, *Age*, *Proposal_delivery* and *Sex*. The *'Indicators'* group of attributes include the attributes derived from the *Loan_amount* column in the analysed data. For the purpose of the case study, it is important to mention the type of the loan – it is a *building savings* loan. Note that in TKR, the *mortgage* loan amount is a measure of TK. Four variants of the *Loan_amount* are created to get the maximum chance to find interesting relationships. All four attributes are included in one class of the equivalence *'Loan'*. It means that in one rule, only one of the four attributes can appear. The attributes differ in the number of categories which are created and the algorithm which is used to create them (*equifrequent*, *equidistant*). The *Equifrequent* option creates categories with the same number of objects (clients). The *Equidistant* option creates categories with the same length of intervals.

- *Loan_ed5*: Equidistant intervals, 5 categories, the class of equivalence *'Loan'*, levels - *very high*, *high*, *medium*, *low*, *very low*, for the respective category
- *Loan_ef5*: Equifrequent intervals, 5 categories, the class of equivalence *'Loan'*, levels - *very high, high, medium, low, very low*, for the respective category
- *Loan_ef11*: equifrequent intervals, 11 categories, levels (overlapping) as shown in Table VII, the class of equivalence *'Loan'*

- *Loan_ed11*: equidistant intervals, 11 categories, levels (overlapping) as shown in Table VII, the class of equivalence *'Loan'*

Moreover, in the task definition, different coefficients are used. For the *Loan_ed5* and *Loan_ef5*, a subset is used – that is, a usual attribute-value pair creation. For the *Loan_ed11* and *Loan_ef11*, a sequence of minimal length 1 and maximal length 3 is used. The inclusion of the derived attributes in the *'Loan'* class of equivalence ensures that only one attribute representing the column Loan from the data matrix will be present in the rule.

The task definition is as follows. In the antecedent, Client properties and the Indicators group of attributes is set. In the consequent, the attribute District is placed. A minimal support is set to 30 objects, minimal Lift=1.5.

C. *Results of the task*

After 3 seconds, 221 rules were found. To obtain less rules which are potentially interesting, it is possible to filter out the consequences of AT/K-formulas contained in TKR. Following the TK Framework of A/TK-formulas (see Fig. 5), TKR is queried to obtain relevant formulas. The *Loan_amount* attribute is present in the resulting rules and the connecting element *District* is present in both resulting rules and A/TK-formulas. This means that the A/TK-formulas *Loan_amount* ↑↑ *Income* and *Loan_amount* ↑↑ *Price of flat* are relevant.

After filtering out the consequences of *Loan_amount* ↑↑ *Income*, 100 rules remain. 121 rules are consequences of the *Loan_amount* ↑↑ *Income*. 24 rules are consequences of the *Loan_amount* ↑↑ *Price of flat*, 20 of them are also consequences of the *Loan_amount* ↑↑ *Income*. This means that 4 additional rules are filtered out and 96 rules remain.

After the consequences of A/TK-formulas are filtered out, we query the TKR for explanations for the remaining 96 rules. 66 of them have a relevant explanation which is based on the measure of the TK *Average amount of mortgage*. It makes no sense to use explanations based on *Income* and *Price of flat*, because those measures of TK were already used in A/TK-formulas.

Table VII: Levels for Loan attributes with 11 categories

Loan category (Loan_ef11, Loan_ed11)	**Rank**	**Levels**
c1	1	Very low
c2	2	Very low
c3	3	Very low / low
…	…	…
c9	9	High / Very high
c10	10	Very high
c11	11	Very high

As for explanations, let us take following rule as an example:

Agent(internal) & Collection(No) & Loan_ed11(ed_8..ed_10)
=> (1.69) District(Hlavni mesto Praha)

The rule says that *percentage of clients having internal agent, not having collections, having high loan and living in the district Hlavni mesto Praha is of 1.69 times higher than the percentage of clients living in the district Hlavni mesto Praha.*

Found explanation for that rule is:

Hlavni mesto Praha => mortgage_avg(very high)

The explanation says that in *the district Hlavni mesto Praha, average amount of mortgage is very high*. This explanation could explain the high amount of the building savings loan (coefficient *c8..c10*, present in the resulting rule) because both loans (*building savings loan* and *mortgage*) are loans for housing purposes which behave very similarly. This knowledge (as stated in the last sentence) is a part of user's domain knowledge; the explanation supports and enhances user's domain knowledge.

Note that the framework of A/TK-formulas is not yet implemented. The features presented above were solved manually.

V. Conclusions and future work

I have defined two approaches to applying Trusted Knowledge – *A/TK-formulas* and the *Explanation system*. As shown in section IV, using A/TK-formulas can significantly reduce the number of resulting patterns which are generated by the data mining system. This helps the user to concentrate on the rules which are interesting from the user's perspective (they are not the consequences of known Trusted Knowledge). Moreover, as introduced in section III.B, the explanations enhancing user's knowledge could help the user to better understand the results of the task.

Furthermore, I have defined a new way of assigning categories to the levels of attributes and measures of TK – the *rank-based approach* based on assignment rules.

In Fig. 7, I summarise the three approaches introduced in the paper – SI-formulas, A/TK-formulas and the Explanation system. I distinguish three types of knowledge. Data knowledge is the knowledge obtained applying the data mining system on the data. Domain knowledge is the knowledge obtained from domain experts. Trusted Knowledge is the knowledge obtained from trusted organisations such as state ministries, statistical offices and so on. Based on this categorisation, it is possible to describe the three approaches mentioned above.

A. *FOFRADAR and SI-formulas*

Domain knowledge in the form of mutual influence of attributes from analysed data (SI-formulas) is used to draw consequences and obtain adjusted (filtered) results of the task. By now, the levels of each attribute were created in coordination with the domain expert. Additionally, a newly proposed *rank-based approach* was introduced to create

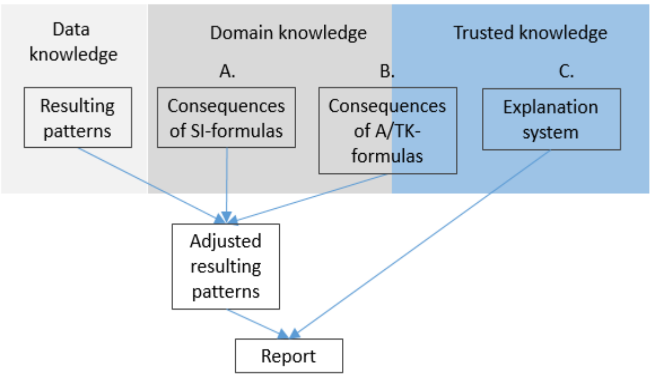

Fig. 7: Summary of used approaches

levels of attributes automatically in situations when the domain expert is not available.

B. A/TK-formulas

A/TK-formulas utilize the newly proposed concept of Trusted Knowledge in combination with the concept of mutual influence introduced in FOFRADAR. This combination of the domain knowledge and Trusted Knowledge is also highlighted in Fig. 7. In this case, it is a mutual influence between an attribute from analysed data and a measure of TK. The levels of the attribute derived from analysed data can be created either in coordination with the domain expert or with the use of the *rank-based approach*. The levels of the measure of TK are created using the *rank-based approach*. The consequences of the mutual influence are drawn automatically and again, adjusted results are obtained. A connecting element is used to connect the measure of TK and resulting patterns.

C. Explanation system

The third approach uses the concept of *explanations*. Trusted Knowledge in the form of the explanations which are relevant to the results of the task is offered to the user. The consequences of Trusted Knowledge are not drawn automatically. The reasoning is left to the user. A connecting element is used to connect an explanation and resulting patterns.

Implementation of A/TK-formulas is left for the future work. The framework also needs to be tested on more complex rules. To get even more benefits from the proposed framework, a sort of publically available sharing of TKR seems to be a next logical step, as for example in [17]. Moreover, a further automation of the task definition and drawing conclusions will be possible due to the defined LMCL scripting language. Another way how to elaborate upon the framework is to applicate it to the data mining with histograms [18].

REFERENCES

[1] Qiang, Y., Xindong, W., 2006. *10 Challenging Problems in Data Mining Research*, International Journal of Information Technology & Decision Making, Vol. 5, No. 4, 2006, 597-604. DOI: 10.1142/S0219622006002258

[2] Mansingh, G., Osei-Bryson, K.-M., Reichgelt. H.: Using ontologies to facilitate post-processing of association rules by domain experts, Information Sciences, 181(3), 2011, 419–434. DOI: 10.1016/j.ins.2010.09.027

[3] Rauch, J., 2015. *Formal Framework for Data Mining with Association Rules and Domain Knowledge – Overview of an Approach*. Fundamenta Informaticae, 137 No 2, pp. 1–47. DOI: 10.3233/FI-2015-1175

[4] Silberschatz, A., Tuzhilin, A., 1995. *On subjective measures of interestingness in knowledge discovery*. In Proc. of the 1st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pages 275-281, 1995. DOI: 10.1.1.88.146

[5] Padmanabhan, B., Tuzhilin, A., 1998. *A belief-driven method for discovering unexpected patterns*. In Proc. of the 4th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pages 94-100, 1998. DOI: 10.1.1.28.728

[6] De Bie, T., 2013. *Subjective interestingness in exploratory data mining*. In Advances in Intelligent Data Analysis XII: 12th International Symposium, IDA 2013, London, UK, October 17-19, 2013. DOI: 10.1007/978-3-642-41398-8_3

[7] Paulheim, H., Ristoski, P., Mitichkin, E., Bizer, C., 2014. *Data Mining with Background Knowledge from the Web*. RapidMiner World, At Boston, USA. August 2014

[8] Paulheim, H., 2012. Generating possible interpretations for statistics from linked open data, in: 9th Extended Semantic Web Conference, ESWC, 2012.

[9] Z. Huang, H. Chen, T. Yu, H. Sheng, Z. Luo, Y. Mao, 2009. Semantic text mining with linked data, in: INC, IMS and IDC, 2009. NCM'09. Fifth International Joint Conference on, 2009, pp. 338–343. DOI: 10.1109/NCM.2009.131

[10] Tiddi I., d'Aquin M., Motta E. 2014. Dedalo: Looking for Clusters Explanations in a Labyrinth of Linked Data. In: Presutti V., d'Amato C., Gandon F., d'Aquin M., Staab S., Tordai A. (eds) The Semantic Web: Trends and Challenges. ESWC 2014. Lecture Notes in Computer Science, vol 8465. Springer, pp. 333-348. DOI: 10.1007/978-3-319-07443-6_23

[11] Czech Statistical Office (CSO), 2015. Výsledky sčítání lidu, domů a bytů 2011 (Census 2011 – in Czech) [online]. https://www.czso.cz/csu/czso/otevrena_data_pro_vysledky_scitani_lidu_domu_a_bytu_2011_-sldb_2011- Last modified on 14 th April 2015.

[12] Buchanan, B. G., Smith, R. G., 1988. *Fundamentals of expert systems*. Annual review of computer science, 1988, 3.1: 23-58.

[13] Rauch, Jan. *Observational Calculi and Association Rules* [online]. 1. ed. Berlin : Springer-Verlag, 2013. ISBN 978-3-642-11736-7. Available at: http://link.springer.com/book/10.1007/978-3-642-11737-4

[14] Šimůnek, Milan. 2014. LISp-Miner Control Language – description of scripting language implementation. Journal of Systems Integration [online], Vol 5, No 2 (2014), p. 28-44. ISSN 1804-2724. URL: http://www.si-journal.org/index.php/JSI/article/view/193 DOI: http://dx.doi.org/10.20470/jsi.v5i2.193

[15] Deloitte Real Index Q3 2016, (in Czech) [online]. Available at https://www2.deloitte.com/content/dam/Deloitte/cz/Documents/real-estate/Deloitte_Real_Index_Q3_2016_CZ.pdf

[16] Czech Ministry of Regional Development. Stav hypotečních úvěrů v krajích za leden až prosinec 2016 (in Czech). Available at http://www.mmr.cz/getmedia/a5bd12f0-2322-4037-80d4-648163c28e50/Stav-hypotecnich-uveru-v-krajich-za-leden-az-prosinec-2016,-s-logem.pdf

[17] Vanschoren, J. 2012. The Experiment Database for Machine Learning (demo) [electronic document]. Workshop PlanLearn 2012. Available from http://datamining.liacs.nl/planlearnpapers/planlearn2012_submission_7.pdf

[18] Rauch, Jan, Šimůnek, Milan. 2015. Data Mining with Histograms – A Case Study. In: *Foundations of Intelligent Systems* [online]. Lyon, 21.10.2015 – 23.10.2015. Cham : Springer International Publishing, 2015, s. 3–8. ISBN 978-3-319-25251-3. DOI: 10.1007/978-3-319-25252-0.