Information Modelling and Knowledge Bases XXXII
M. Tropmann-Frick et al. (Eds.)

doi:10.3233/FAIA200832

Search for Appropriate Textual Information Sources

Adam ALBERT[1], Marie DUŽÍ[1], Marek MENŠÍK[1], Miroslav PAJR[2], Vojtěch PATSCHKA[1]
[1] VSB-Technical University Ostrava, Department of Computer Science FEI
17. listopadu 15, 708 33 Ostrava, Czech Republic
[2] Silesian University in Opava, Institute of Computer Science,
Bezručovo nám. 13, 746 01 Opava, Czech Republic

Abstract. In this paper, we deal with the support in the search for appropriate textual sources. Users ask for an atomic concept that is explicated using machine learning methods applied to different textual sources. Next, we deal with the so-obtained explications to provide even more useful information. To this end, we apply the method of computing association rules. The method is one of the data-mining methods used for information retrieval. Our background theory is the system of Transparent Intensional Logic (TIL); all the concepts are formalised as TIL constructions.

Keywords. Machine learning; Transparent Intensional Logic; TIL; atomic concept; molecular concept; association rules; explication; natural language processing; information source recommendation

1. Introduction

We live in the era of globalisation, i.e. in times of interaction among people worldwide that has grown due to great advances in transportation, information and communication technologies. Though being a complex phenomenon, globalisation is usually characterised as a form of the integration of local economies into a global, unregulated market economy. Yet, the same phenomenon can be traced in other spheres of our lives, including science and research development. Globalisation has positive effects on the environment, culture, economic development, and in general human well-being in societies around the world. These are the upsides. However, some people complain that there are the downsides of globalisation as well because the gaps between rich and developing countries have grown. In 2000, the International Monetary Fund (IMF) identified four essential aspects of globalisation: trade and transactions, capital and investment movements, migration and movement of people, and the dissemination of knowledge.[1]

Here we are not going to deal with economic aspects of this phenomenon; rather, we are interested in the last issue mentioned, namely the increasing amount of knowledge, technology and information moving across international borders. Together with human innovation and progress in information technologies, these factors give rise to '*information overload*'. True, there is a lot of knowledge around, spread in the vast amount of textual resources available on the Internet. Yet, there are also plenty of disinformation,

[1] For details, see [7].

fake news, futile texts, canards, merely a lot of words, words, words. Many of our fellow researches certainly experienced the situation of googling for relevant, high-quality papers from reliable resources, only to the effect of obtaining a lot of futile results. The gist of the story is this. There is a need for an 'intelligent' question-answering system that would not only search by keywords but also evaluate the results, check for inconsistencies, derive their consequences logically entailed or just semantically associated with them, etc.

To this end, we decided to start a project on building such a system. Our background theory is the system of Transparent Intensional Logic (TIL) that makes it possible to formalise all the semantically salient features of natural language in a fine-grained way.[2] Having formalised thousands of sentences in the form of TIL-constructions, we can now proceed in two closely interrelated ways.[3] First, we have integrated special rules rooted in the rich semantics of natural language into a standard proof calculus of Genzen's natural deduction or general resolution in order to derive logical consequences of the results of a prior search.[4] Second, by applying machine-learning methods adjusted to natural language processing in TIL, we *explicate* atomic concepts in order to 'understand' and manipulate them in the way human agents would do if only this task were not beyond their capacities.[5] Here we mean Carnapian *explication* (see [2, pp. 7-8, §2]) that is transformation of a given more or less inexact concept (the explicandum) into an exact one (the explicatum). Our explicandum is usually an atomic concept such as 'cat', 'dog', 'lattice', 'group' to which a molecular concept is assigned that ontologically defines the objects falling under the explicandum. For instance, the concept 'cat' can be explicated by the biological definition of cats like 'a small domesticated carnivorous mammal of the family Felidae'. Having such a molecular concept, we can derive even more useful information from the vast number of textual corpora.

Explications of atomic concepts extracted from one textual document have been introduced in [12]. By applying a supervised machine-learning method to multiple textual resources, we obtain several different explications (see [10]). For instance, by applying the method to the atomic concept 'lattice', we obtain molecular concepts like 'a structure of crossed wooden or metal strips arranged to form a diagonal pattern of open spaces between the strips', 'a window, gate, or the like consisting of such a structure', or in physics 'the structure of fissionable and non-fissionable materials geometrically arranged within a nuclear reactor'. In mathematics, we can find two equivalent definitions of an abstract lattice structure, namely 'a partially ordered set in which every subset of two elements has a unique supremum and infimum' or 'an algebra with two binary operations meet and join that satisfy the axioms of commutativity, associativity and absorption'. In music, the same atomic concept can mean 'an organised grid model of pitch ratios.'

The question now arises how to evaluate these results so that to be able to recommend the most relevant, appropriate resources. There are several possibilities. We can check the results for inconsistencies or similarities, extract or generalise what they have

[2] See, for instance, [17].

[3] The text data are linguistically and logically processed so that TIL constructions are extracted by the Natural Language Logical Analyzer algorithm, see [9].

[4] For details, see [4], [5].

[5] The first application checking the possibilities of supervised machine learning methods adjusted to natural language processing has been introduced in [11] where the method has been applied to recognition of geometric figures.

in common, etc. The goal of this paper is to introduce the algorithm based on *associations* between the data so that to compute and recommend the most relevant textual resources. For instance, concepts that are *semantically* associated with the above concept of the lattice are 'network', 'web', 'grid', 'structure', 'algebra', 'ordered set'. Yet, we are also interested in associations of concepts that follow from frequencies of their co-occurrence.

Depending on the amount of input data, users can obtain a huge number of different molecular concepts corresponding to the atomic concept that has been asked for. Now, the user can pick up one explication and the corresponding resource that seems to be relevant, but there can be other similar resources that are appropriate as well. Yet, due to a large amount of input textual data, the other proper documents can be overlooked and ignored. Thus, to prevent such a situation, we apply the method of discovering 'hidden' associations between the constituents of the resulting molecular concepts.

The whole process can be described as follows. First, a supervised machine learning method adjusted to TIL is applied to extract molecular concepts explicating the atomic concept that is the subject of the initiative query. As a result, we obtain several such explications that should be further evaluated. To this end, we apply the method of association rules. The system organises constituents of the obtained molecular concepts into an *incident matrix.* The rows of the matrix represent particular explications, and the columns represent the concepts of properties mentioned in those explications. We use a two-valued Boolean matrix. Next step consists in extracting *association rules*. Each rule is of form $A \Rightarrow B$ where A is the antecedent and B the succedent of the rule, and A, B are sets of concepts. The sense of such a rule is this. If a given explication contains all the concepts from A then it is to a certain degree probable that it also contains concepts from B. The so-called *minimal confidence* of an association rule, i.e. conditional probability of occurrences of concepts from B provided there are concepts from A is defined by a user. By computing the rules that are valid at least with this user-defined minimal confidence, the algorithm then proposes other textual sources that might be relevant as well.

The rest of the paper is organised as follows. Section 2 briefly summarises the main principles of Transparent Intensional Logic and introduces TIL constructions that serve as a concept defining formalism. In Section 3, we summarise previous results on the topic, namely the system of seeking relevant textual resources as presented in [12]. In Section 4, we introduce the theory of association rules, while in Section 5, the whole algorithm of their computing is described. Concluding remarks and further research proposals can be found in Section 6.

2. Foundations of TIL

In TIL, expressions encode *algorithmically structured procedures* as their meanings. These procedures produce extensional or intensional entities, or even lower-order procedures, as their products. This approach has summarised by an *algorithmic turn* in semantics and advocated, for instance, by Moschovakis in [14]. Yet much earlier, in the early 1970s, Pavel Tichý defined six kinds of such meaning procedures that he coined TIL *constructions* as the centre-piece of his system, see [6] or [17].

The syntax of TIL is a hyperintensional, typed λ-calculus of partial functions. However, TIL λ-terms do not denote functions; rather, they denote *procedures* (*constructions* in TIL terminology) that produce functions or functional values as their products. A linguistic sense of an expression is an abstract *procedure* detailing how to arrive at an object (if any) of a particular logical type denoted by the expression.

There are two kinds of TIL constructions, atomic and molecular. *Atomic* constructions (*Variables* and *Trivialisations*) do not contain any other constituent but itself; they supply objects (of any type) on which compound constructions operate. Variables *x, y, p, q, …* construct objects dependently on a valuation; they *v*-construct. To each type, countably many variables are assigned, which *v*-constructs elements of the assigned type; we also say that variables range over that type. *Trivialisation* of an object *X* (of any type, even a construction), in TIL symbolism *'X*, refers to or displays the object *X* without the mediation of any other construction. In order to operate on *X*, the object must be grabbed first; Trivialisation is such a simple grabbing mechanism.

There are two dual molecular constructions, namely *Composition* and *Closure*. *Composition* $[F\, A_1 \ldots A_n]$ is the procedure of applying a function *f* (*v*-constructed by the first constituent *F*) to a tuple argument *a* (*v*-constructed by the constituents $A_1, \ldots, A_n$). Composition *v*-constructs the value of *f* at *a*, if the function *f* is defined at *a*, otherwise the *Composition* is *v*-improper, i.e., it fails to *v*-construct anything. To produce a function rather than its value, there is (λ-)*Closure* $[\lambda x_1 \ldots \lambda x_n X]$. It is a procedure of *v*-constructing a function by abstracting over the values of variables $x_1, \ldots, x_n$ in the ordinary manner of λ-calculi. Finally, higher-order constructions producing lower-order constructions can be executed twice over. This is achieved by a fifth construction called *Double Execution*, ^2X, that behaves as follows: If *X v*-constructs a construction *Y*, and *Y v*-constructs an entity *Z*, then ^2X *v*-constructs *Z*; otherwise ^2X is *v*-improper by failing to produce anything.

TIL constructions, as well as the entities they construct, all receive a type within a *ramified hierarchy of types*. Thus, the formal ontology of TIL is *bidimensional*; one dimension is made up of constructions of order $n \geq$!, the other dimension encompasses non-constructions. On the ground level of the type-hierarchy, there are non-procedural entities unstructured from the algorithmic point of view belonging to a *type of order 1*. Given a so-called *epistemic* (or *'objectual'*) base of atomic types (ο-truth values, ι-individuals, τ-time moments/real numbers, ω-possible worlds), the induction rule for forming functions is applied: where α, $\beta_1, \ldots, \beta_n$ are types of order 1, the set of partial mappings from β_1 x ... x β_n to α, denoted $(\alpha\, \beta_1 \ldots \beta_n)$, is a *type of order 1* as well.[6] Constructions that construct entities of a type of order 1 are *constructions of order 1*. They belong to a *type of order 2*, denoted by $*_1$. This type $*_1$ together with atomic types of order 1 serves as a base for the induction rule of forming functions: any collection of partial mappings, type $(\alpha\, \beta_1 \ldots \beta_n)$, involving $*_1$ in their domain or range is a *type of order 2*. Constructions that construct entities belonging to a type of order 1 or 2 are *constructions of order 2*. They belong to a *type of order 3*, denoted $*_2$; any collection of partial mapping involving $*_2$ in their domain or range is a *type of order 3*. And so on ad infinitum.

Empirical sentences and terms denote (PWS-)*intensions*, functions with the domain of possible worlds ω; they are frequently mappings from ω to *chronologies* of α-objects, hence functions of types ((ατ)ω), or $\alpha_{\tau\omega}$, for short. Where variables *w, t* range over possible worlds ($w \rightarrow \omega$) and times ($t \rightarrow \tau$), respectively, constructions of intensions are usually Closures of the form $\lambda w \lambda t\, [\ldots w \ldots t \ldots]$. We model *sets* and *relations* by their

[6] The above epistemic base {ο, ι , τ, ω} was chosen, because it is apt for natural-language analysis, but the choice of the base depends on the area to be analysed.

characteristic functions. Hence, while (οι), (οιι) are types of a set of individuals and of a binary relation-in-extension between individuals, respectively, $(\omicron\iota)_{\tau\omega}$, $(\omicron\iota\iota)_{\tau\omega}$ are types of a *property* of individuals and a binary relation-in-intension between individuals, respectively. *Quantifiers* $\forall^{\alpha}$, $\exists^{\alpha}$ are type-theoretically polymorphic total functions of types (ο(οα)) defined as follows. Where B is a construction that v-constructs a set of α-objects, $[{}^{0}\forall^{\alpha}B]$ v-constructs **T** if B v-constructs the set of all α-objects, otherwise **F**; $[{}^{0}\exists^{\alpha}B]$ v-constructs **T** if B v-constructs a non-empty set, otherwise **F**.

Notational conventions. That an object X belongs to a type α is denoted as 'X/α'; that a construction C v-constructs an α-object (provided not v-improper) is denoted by '$C \rightarrow \alpha$'. Instead of $[‘\forall^{\alpha}\,\lambda x\,A]$, $[‘\exists^{\alpha}\,\lambda x\,A]$ we write '$\forall x\,A$', '$\exists x\,A$' whenever no confusion arises. If $C \rightarrow \alpha_{\tau\omega}$ then the frequently used Composition $[[C\,w]\,t]$, aka extensionalization of the α-intension v-constructed by C, is abbreviated as C_{wt}. We use classical infix notation without Trivialization for truth-value functions ∧ (conjunction), ∨ (disjunction), ⊃ (implication) and ¬ (negation). Also, identities $=^{\alpha}$ of α-objects are written in the infix way without Trivialisation and the superscript α whenever no confusion arises.

Concepts are modelled as *closed constructions* in their normal form. *The atomic concept* of an object a is its Trivialisation, $‘a$, while *molecular concept* of an object a is its ontological definition, i.e. a closed molecular construction producing a. Unlike Frege and in compliance with Church, we deal with concepts of entities of any type, including concepts of propositions of type $\omicron_{\tau\omega}$ in an empirical vernacular and of truth-values in mathematics.[7]

For a simple example, where $Student/(\omicron\iota)_{\tau\omega}$ is a property of individuals and $John/\iota$ an individual, the sentence "John is a student" encodes as its meaning the concept of the proposition

$$\lambda w \lambda t\ [‘Student_{wt}\ ‘John]$$

The property *Student* must be extensionalized first, $‘Student_{wt} \rightarrow (\omicron\iota)$ and only then can it be applied to John, $[‘Student_{wt}\ ‘John] \rightarrow \omicron$, to obtain a truth value according as John belongs to the population of students in a world w and time t of evaluation. Abstracting over the values of variables w, t the proposition of type $\omicron_{\tau\omega}$ that John is a student is produced. The atomic concept $‘Student$ of the property of being a student can be further explicated by a molecular concept, for instance of the property of being a person who attends a school.

$$\lambda w \lambda w\ \lambda x\ [[‘Person_{wt}\ x] \wedge [‘Attend_{wt}\ x\ ‘School]]$$

This completes our brief introduction to the system of TIL and its theory of concepts.

3. Explication of atomic concepts by machine learning

Supervised machine learning is a method of predicting functional dependencies between input values and the output value. The supervisor provides an agent/learner with a set of training data. These data describe an object by a set of attribute values such that there is a functional dependency between these values.[8]

[7] For details on TIL theory of concepts see [6, § 2.2].

[8] In this section we briefly recapitulate the results as presented in [12].

For instance, a house can be characterised by its size, locality, date of being built up, architecture style, etc., and its price. Obviously, the price of a house depends on its size, locality, date of building and architecture style. Hence, the price is called an output attribute, and the other attributes are input attributes. The goal of learning is to discover this functional dependency on the grounds of training data examples so that the agent can predict the value of the output attribute given the values of input attributes of a new instance.[9]

In our project of natural language processing and question answering, we decided to apply this method to learning *concepts*. To this end, we have to adjust the method a bit. First, instead of input/output attributes, we deal with concepts, that is closed constructions. The role of input attributes is played by the constituents of a hypothetic molecular concept, and instead of the output attribute, we deal with the atomic concept that the learner wants to learn by refining examples extracted from the textual documents. The hypothetic function is the relation of a requisite, or typical property or even a semantic association. Training data are natural-language texts, and the supervisor extracts from the text data positive and negative examples. The general framework of machine learning based on symbolic representation consists of the learning objectives, training data and heuristic methods for manipulating the symbolic representation of the data. For our purpose, we voted for an adjusted version of Patrick Winston algorithm [18, pp. 349-363] of supervised machine learning. This algorithm applies the principles of *generalisation* and *specialisation* to obtain a plausible hypothesis. Another adjustment of the algorithm is this. In addition to generalisation and specialisation, we also use the method of *refinement*. By refining a hypothetic concept, we insert new constituents into the molecular construction learned so far.

Generalisation usually consists in replacing one or more constituents of the hypothetic concept by a more general one, which is either extracted from agent's ontology or created from the chosen constituents by composing them in a disjunctive way. As a particular case, generalisation can also be applied to numerical values of attributes. For instance, if we obtain a piece of information that the in-heat period of a wild cat is two days and another positive example specifies eight days, we generalise it to the interval 2 – 8 days.

Specialisation is triggered by negative examples. As a result, the negation of a property that does not belong to the essence of the hypothetic concept is inserted. Specialisation serves to distinguish the concept from similar ones. For instance, a wooden horse can serve as a negative example to the concept of a horse because a wooden horse is not a horse; rather, it is a toy horse though it may look like a genuine living horse.

For example, let the 'output' concept (to be learned) be that of a cat, i.e. *'Cat.* The role of positive examples is played by ontological definitions of the property of being a cat, like "Cat is a predatory mammal that has been domesticated". The learner establishes a hypothesis that the property

$$\lambda w \lambda t\, \lambda x\, [[[{}'Predatory\ {}'Mammal]_{wt}\, x] \wedge [{}'Domesticated_{wt}\, x]]$$

belongs to the essence of the property *Cat.* Negative examples delineate the hypothesis from other similar objects. For instance, the sentence "Dog is a domesticated predatory mammal that barks" can serve as a negative example for *Cat.* This triggers a specialisation of the hypothetic concept to the construction

[9] For details, see [13, pp. 20-28], [15, Ch. 7.8], [16, pp. 695-697].

$$\lambda w \lambda t\ \lambda x\ [[[\text{'}Predatory\ \text{'}Mammal]_{wt}\ x] \wedge [\text{'}Domesticated_{wt}\ x] \wedge \neg[\text{'}Bark_{wt}\ x] \wedge \neg[\text{'}Dog_{wt}\ x]]$$

Hence, given a positive example, the learner *refines* the hypothetic molecular concept by adding other constituents to the essence, while a negative example triggers *specialisation* of the hypotheses. The hypothetic concept can also be *generalised*. For instance, the learner can obtain the sentence “Cat is a *wild* feline predatory mammal” as another positive example describing the property *Cat*. Since the properties *Wild* and *Domesticated* are inconsistent, the agent consults his/her ontology for a more general concept. If there is none, the ‘union’ of the properties, *Wild* or *Domesticated*, is included. As a result, the learner obtains this hypothesis.

$$\lambda w \lambda t\ \lambda x\ [[[\text{'}Feline\ [\text{'}Predatory\ \text{'}Mammal]]_{wt}\ x] \wedge [[\text{'}Domesticated_{wt}\ x] \vee [\text{'}Wild_{wt}\ x]] \wedge \neg[\text{'}Bark_{wt}\ x] \wedge \neg[\text{'}Dog_{wt}\ x]]$$

Remark. Both *Feline* and *Predatory* are property modifiers of type $((o\iota)_{\tau\omega}(o\iota)_{\tau\omega})$, i.e. functions that given a root property return another property as an output. Since these two modifiers are intersective, the rules of left- and right-subsectivity are applicable here.[10] In other words, the predatory mammal is a predator and is a mammal, similarly for a feline. If our agents have these rules in their knowledge base, the above Composition

$$[[\text{'}Feline\ [\text{'}Predatory\ \text{'}Mammal]]_{wt}\ x]$$

can be further refined to

$$[[\text{'}Feline^{p}{}_{wt}\, x] \wedge [\text{'}Predator^{p}{}_{wt}\, x] \wedge [\text{'}Mammal_{wt}\ x]],$$

where $Feline^{p}$ and $Predator^{p}$ are properties of individuals, i.e. objects of type $(o\iota)_{\tau\omega}$.

Heuristic methods of the original Winston algorithm work with examples that cover all the attributes of a learned object. Based on positive examples, the hypothesis is modified in such a way that the values of attributes are adjusted, or in case of a negative example, an unwanted attribute marked as Must-not-be is inserted. In our application, the sentences that mention the learned concept contain as constituents some but not all the requisites of this concept, and we build up a new molecular concept by adding new information extracted from positive or negative examples. Hence, we had to implement a new heuristic *Concept-introduction* for adding concepts of new requisites into a hypothetic concept. Negative examples trigger the method *Negative-concept* that inserts a concept of negated property into the hypothesis. Generalisation is realised by modules that introduce a concept of a more general property; to this end, we also adjusted the original heuristic *Close-interval* so that it is possible to generalise values of numeric concepts by the union of interval values.

Description of the Explication algorithm

Here is a brief specification of the algorithm.

Refinement.

1. Compare the model hypothesis (to be refined) and the positive example to find a significant difference
2. If there is a significant difference, then

[10] For details on and analysis of modifiers, see [3]. Details on the way of integrating such special semantic rules into a standard proof calculus can be found in [5], [4].

```
    a) if the positive example contains as its constituent a concept that the model
       does not have, use the Concept-introduction
    b) else ignore example
```

Specialisation.

```
1. Compare the model hypothesis (to be refined) and the near-miss negative example
   to find a significant difference
2. If there is a significant difference, then
    c) if the near-miss example has a constituent of the concept that the model does
       not have, use the Negative-concept
    d) else ignore example
```

Generalisation.

```
1. Compare the model hypothesis (to be refined) and the positive example to determine
   a difference
2. For each difference do
    a) if a concept in the model points at a value that differs from the value in the
       example, then
        i)  if the properties in which the model and example differ have the most
            specific general property, use the General-concept
        ii) else use Disjunctive-concept
    b) if the model and example differ at an attribute numerical value or interval, use
       the Close-interval
    c) else ignore example.
```

4. Association rules

The method of *association rules* extraction has been introduced in [1]. Yet ten years earlier a similar method has been described in [8]. Basically, it is the process of looking for interesting relations among a large number of items. The method can be applied in various areas such as market survey or risk management, and a typical application is a market basket analysis. The goal is to discover associations between items occurring in a dataset that satisfy predefined minimum support and confidence.

The algorithm first extracts frequent item-sets, i.e. those item-sets whose occurrences exceed a predefined threshold k called minimal *support*. Then *confidence* of associations among these frequent item-sets is computed and compared with predefined minimal *confidence*. Only those associations that exceed the predefined minimal support and confidence are then considered to be interesting results of the data mining method.

To put these ideas on more solid ground, here are the definitions. First, we need to define the *support* of a given set A of items in a dataset. It is the probability of an occurrence of the set A in the entire dataset.

Definition (*support*). Let $I = \{i_1, \ldots, i_n\}$ be a set of items and $D = \{T_1, \ldots, T_m\}$ a dataset of records such that each $T_\mathrm{i} \subseteq I$. Then *support* of a set of items $A \subseteq I$ in D is

$$supp(A) = \frac{|\{t \in D: A \subseteq t\}|}{|D|}$$

Remark. By $|S|$ we denote the cardinality of a set S. Since $|D| = m$, the support of a set A is the ratio that compares the number of records containing all items from A to the total number m of records in the dataset. Hence, the support of A is the probability of the occurrence of items from A in the dataset.

Definition (association *rule, confidence*). Let $I = \{i_1, \ldots, i_n\}$ be a set of items and $D = \{T_1, \ldots, T_m\}$ a dataset of records such that each $T_i \subseteq I$. Farther, let $A, B \subseteq I$ such that $supp(A \cup B) \geq k$, where k is a predefined threshold. Then $A \Rightarrow B$ is an *association rule* iff $A \cap B = \emptyset$ and $A, B \neq \emptyset$. *Confidence* of the rule $A \Rightarrow B$ is

$$conf(A \Longrightarrow B) = \frac{supp(A \cup B)}{supp(A)}$$

Example. Let us have the following *dataset* of shopping transactions:

D = {T1, T2, T3, T4, T5}
T_1 = {milk, butter},
T_2 = {bread, milk, butter},
T_3 = {milk},
T_4 = {bread, milk, butter},
T_5 = {bread, butter}

Then the incident matrix is this.

D	bread	milk	butter
T_1	0	1	1
T_2	1	1	1
T_3	0	1	0
T_4	1	1	1
T_5	1	0	1

Let *min-supp* = 0.25 and *min-conf* = 0.75. Then there are the following association rules meeting the thresholds:

supp ({milk, butter}) = 3/5
conf ({milk}⇒{butter}) = 3/4
conf ({butter}⇒{milk}) = 3/4

supp ({bread, butter}) = 3/5
conf ({bread}⇒{butter}) = 1
conf ({butter}⇒{bread}) = 3/4

supp ({bread, milk, butter}) = 2/5
conf ({bread, milk}⇒{butter}) = 1

As the example illustrates, the method can be applied for instance, in e-shops to recommend other products to be bought once a customer inserts into the shopping basket a given set of products. This feature inspired us to apply the method in our system in order to recommend other possible interesting explications of a given concept once a user votes for one of the obtained explications.

5. Algorithm of text-source recommendations

In this chapter, we summarise the whole system including the modules of supervised machine learning introduced in [10]. For the system outline see Fig. 1; the new functionality is incorporated in the very last part of the system, namely the *Relevant Source Selection* module.

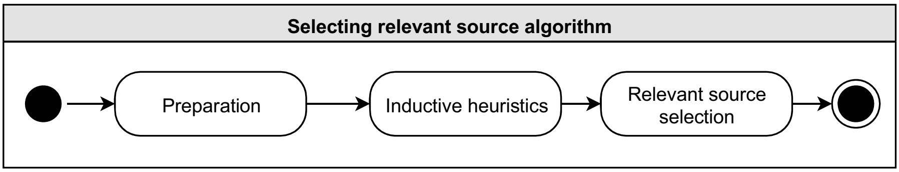
Selecting relevant source algorithm
Preparation
Inductive heuristics
Relevant source selection

Fig. 1. Algorithm outline

First, we need to analyse textual resources to obtain the base of formalised TIL constructions. To this end, linguistic and logical analysis is applied.[11] Then the set of relevant propositional constructions is selected; namely those where the concept to be explicated occurs.

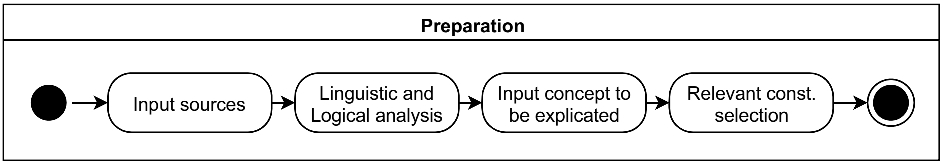
Preparation
Input sources
Linguistic and Logical analysis
Input concept to be explicated
Relevant const. selection

Fig. 2. Pre-processing and formalisation of textual resources

Next, the set of the selected constructions serves as an input for machine learning techniques, in particular, the *Inductive heuristics* module (see [12]), to obtain plausible hypotheses that explicate the given simple concept (Fig. 3). In this way, we obtain several explications, each of which corresponds to one of the input textual documents.

[11] Textual data are linguistically and logically processed so that TIL constructions are extracted by the Natural Language Logical Analyzer algorithm [9].

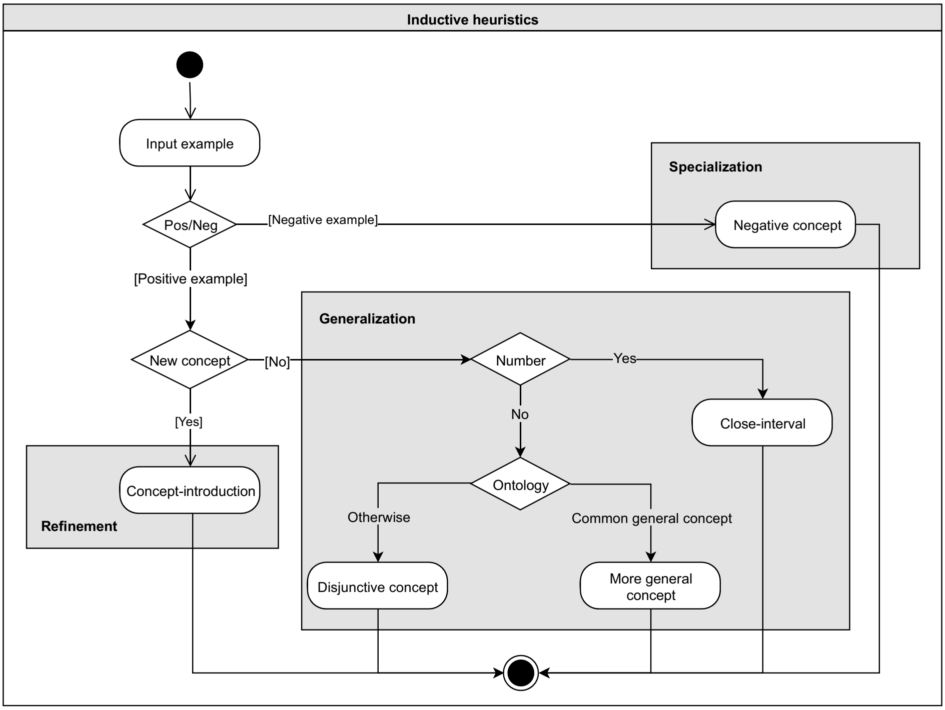

Fig. 3. Supervised machine learning

The last module of *Relevant Source Selection* (Fig. 4) is still work in progress. It is the module that deals with hypotheses processing and their evaluation. There are several functionalities that might be realised here. They include, inter alia, filtering out irrelevant sources according to the additional user-defined linguistic and logical criteria, search for inconsistencies among the hypotheses such as *contrarieties* and *contradictions*, looking for striking news that defies our intuitions and as such might be fake news, checking the reliability of resources.

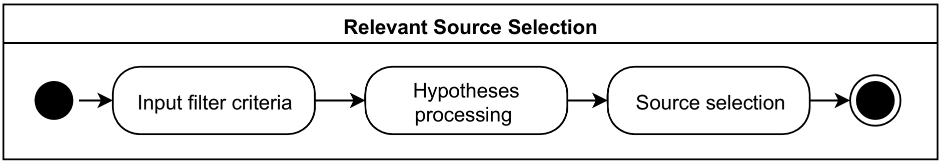

Fig. 4. Relevant source selection

Our *recommendation system* introduced in this paper is incorporated in the module *Hypotheses processing*, see Fig. 5.

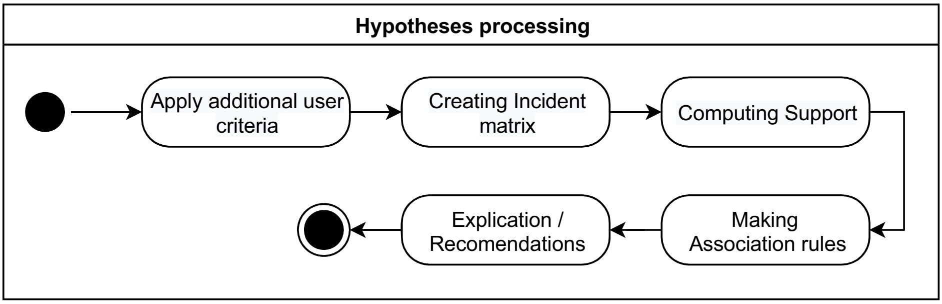

Fig. 5. Hypotheses processing

The input data to this module are a collection of closed constructions, i.e. the molecular concepts extracted from textual documents by the machine-learning module as explications of the simple concept *c* asked for. The algorithm keeps the track of the document from which particular explications have been extracted. The algorithm extracts from each of these explications *e* the concepts of properties $P_j \rightarrow (o\alpha)_{\tau\omega}$ such that these P_j are 'conjunctive constituents' of *e*. By 'conjunctive constituents' we mean those subconstructions that are connected in a conjunctive way. For instance, the conjunctive constituents of the molecular concept

$$\lambda w \lambda t\ \lambda x\ [[[\text{'}Feline\ [\text{'}Predatory\ \text{'}Mammal]]_{wt}\ x] \wedge$$
$$[[\text{'}Domesticated_{wt}\ x] \vee [\text{'}Wild_{wt}\ x]] \wedge$$
$$\neg[\text{'}Bark_{wt}\ x] \wedge \neg[\text{'}Dog_{wt}\ x]]$$

are these:

$\lambda w \lambda t\ \lambda x\ [[\text{'}Feline\ [\text{'}Predatory\ \text{'}Mammal]]_{wt}\ x]$,
$\lambda w \lambda t\ \lambda x\ [[\text{'}Domesticated_{wt}\ x] \vee [\text{'}Wild_{wt}\ x]]$,
$\lambda w \lambda t\ \lambda x\ \neg[\text{'}Bark_{wt}\ x]$,
$\lambda w \lambda t\ \lambda x\ \neg[\text{'}Dog_{wt}\ x]$

Let *I* be the set of all the constituents extracted from the explications. The algorithm now computes binary *incident matrix*, rows of which represent explications and items of columns are constituents $i \in I$. Farther, association rules are computed from this incident matrix. Recall that association rule $A \Rightarrow B$ represents an association between disjoint sets of sufficiently frequent non-empty sets of items in a given dataset. Our dataset is now a set of records (rows of incident matrix) extracted from particular explications, see Table 1 below.

The user selects one of the input explications that is the closest one to his/her intuitive idea explaining the simple concept *c* asked for. The goal is to find other explications (and thus text documents as well) which concern the concept *c* and might be potentially interesting for the user. Association rules that serve as those recommending other explications are computed with respect to these criteria.

1. Antecedent contains only those constituents which occur in the selected explication *e*.
2. Succedent contains only the remaining constituents from *I* which do not occur in the selected explication *e*.
3. Support and confidence of the rule are greater or equal to the value of the predefined criteria *min-supp* and *min-conf*.

Formally, these criteria are defined as follows.

Definition: Let $A \Rightarrow B$ be an association rule, $E=\{e_1,\ldots,e_n\}$ the set of all explications, $e \in E$ the user-selected explication, and let *Prop*(x) be the set of all constituents occurring in an explication x. Then the *rule* $A \Rightarrow_e B$ is a *rule of recommendation* generated by the selected explication e iff:

$$A \subseteq Prop(e)$$

$$B \subseteq \left(\bigcup_{i=1}^{n} Prop(e_i) \right) \setminus Prop(e)$$

$$supp(A \cup B) \geq min\text{-}supp$$

$$conf(A \Rightarrow B) \geq min\text{-}conf$$

Remark. Obviously, to each explication e there can be more than one rule of recommendation generated by e.

Having computed the rules of recommendation, we want to recommend other documents dealing with the input concepts c. Thus, we define:

Definition: Let $A \Rightarrow_e B$ be a *rule of recommendation generated by* the selected explication e. Let *exp(d,c)* be an explication of an input simple concept c extracted from a textual document d. Then the *recommended sources* dealing with the concept c according to the rule $A \Rightarrow_e B$ is a set of text-sources *RS* such that

$$RS = \{d\colon (A \cup B) \subseteq Prop(exp(d,c))\}$$

Moreover, *weakly recommended sources* explaining the concept c is a set of text-sources *WRS* such that

$$WRS = \{d\colon B \subseteq Prop(exp(d,c))\}$$

5.1. Case study example

In our case study, we had eight documents, i.e. text-sources, dealing with the concept of a wild cat. From each document, the algorithm selected those sentences where ‘wild cat’ receives mention. These sentences have been formalised as TIL constructions explicating the concept ‘*wild cat*’.

Remark. In the constructions below, we use two relations between properties, namely *Req* (for a requisite) and *Typ-p* (typical property). Though the differentiation between *Req* and *Typ-p* is irrelevant for the purposes of this paper, we briefly explain. The first one obtains between two properties P and Q necessarily. Hence, [*Req P Q*], i.e. P is a requisite of Q, should be understood like this: necessarily, if an individual a happens to be a Q then a is a P. On the other hand, [*Typ-p P Q*], i.e. P is typical for Q, is to be read as follows: Typically, if an individual a happens to be a Q then a is a P. Note that both these sentences should be read *de dicto.* They talk about *properties* (intensions) rather than about a particular individual. Hence, that having a fur is a requisite of the property of being a wild cat does not exclude the possibility that this or that cat lost its fur.

Source 1.

- The weight of a wild cat is between 1.2 and 11 kilograms.
- Wild cats are mammals.
- Wild cats have fur.
- The body length of wild cats is from 47 to 80 cm.
- The average skull capacity of wild cats is 41.25 cm^3.
- The average height of wild cats at the withers is 37.6 cm.

Exp (Source 1, ‘*Wild-cat*).

$$\big[‘\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \big[[‘\leq [‘\mathit{Weight}_{wt}x]\ ‘11] \wedge [‘\geq [‘\mathit{Weight}_{wt}x]\ ‘1.2]]\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\big] \wedge$$

$$\big[‘\mathit{Req}\ ‘\mathit{Mammal}\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\big] \wedge \big[‘\mathit{Req}\ ‘\mathit{Has\text{-}fur}\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\big] \wedge$$

$$\big[‘\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \big[[‘\leq [[‘\mathit{Average}\ ‘\mathit{Body\text{-}Length}]_{\mathrm{wt}}x]\ ‘80] \\ \wedge [‘\geq [[‘\mathit{Average}\ ‘\mathit{Body\text{-}Length}]_{wt}x]\ ‘47]\big]\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\big] \wedge$$

$$\big[‘\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \big[‘= [[‘\mathit{Average}\ ‘\mathit{Skul\text{-}Size}]_{\mathrm{wt}}x]\ ’41.25\big]\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\big] \wedge$$

$$\big[‘\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \big[‘= [[‘\mathit{Average}\ ‘\mathit{Height}]_{\mathrm{wt}}x]\ ’37.6\big]\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\big]$$

Source 2.

- The typical occurrence of wild cats is mixed or deciduous forests.
- The size of the territory of a wild cat is greater than 50 ha.
- Wild cat marks its territory with its claws, urination and droppings.

Exp (Source 2, ‘*Wild-cat*).

$$\big[‘\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \big[‘\mathit{Live\text{-}in}_{wt}\ x\ \lambda w\lambda t\ \lambda y\ [[[‘\mathit{Mixed}\ ‘\mathit{Forest}]_{wt}y] \\ \vee [[‘\mathit{Decidious}\ ‘\mathit{Forest}]_{wt}y]]\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\big] \wedge$$

$$\big[‘\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \big[‘\geq [‘\mathit{Territory\text{-}Size}_{wt}x]\ ‘47\big]\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\big] \wedge$$

$$\big[‘\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \big[[‘\mathit{Ter\text{-}Marking}_{wt}x\ ‘\mathit{Clawing}] \vee [‘\mathit{Ter\text{-}Marking}_{wt}x\ ‘\mathit{Urinating}] \vee \\ [‘\mathit{Ter\text{-}Marking}_{wt}x\ ‘\mathit{Droppings}]\big]\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\big]$$

Source 3.

- The in-heat period of the wild cat is 2 – 8 days.
- The wild cat is looking for a mate with a loud meow.
- The pregnancy period of a wild cat is 65 days.
- The size of the litter of wild cats is 3 – 4 kittens.

Exp (Source 3, ‘*Wild-cat*).

$$\big[‘\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \big[[‘\leq [‘\mathit{In\text{-}Heat\text{-}Period}_{wt}\,x]\ ‘8] \wedge [‘\geq [‘\mathit{In\text{-}Heat\text{-}Period}_{wt}\,x]\ ‘2]\big]\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\big] \wedge$$

$$\big[‘\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \big[‘\mathit{Seek}_{wt}\,x\ ‘\mathit{Mate}\ [‘\mathit{Loud}\ ‘\mathit{Meow}]\big]\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\big] \wedge$$

$$\big[‘\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \big[‘= [‘\mathit{Pregnancy\text{-}Period}_{wt}\,x]\ ‘65\big]\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\big] \wedge$$

$$\big[‘\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \big[[‘\leq [‘\mathit{Litter\text{-}Size}_{wt}\,x]\ ‘4] \wedge [‘\geq [‘\mathit{Litter\text{-}Size}_{wt}\,x]\ ‘3]\big]\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\big]$$

Source 4.

- Wild cats are mammals.
- Wild cats have fur.
- The average skull capacity of a wild cat is 41.25 cm^3.
- Wild cats mark their territory with claws, urination, droppings.
- The pregnancy period of a wild cat is 65 days.
- The size of the litter of wild cats is up to 4 kittens.

Exp (Source 4, ‘*Wild-cat*).

$$\big[‘\mathit{Req}\ ‘\mathit{Mammal}\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\big] \wedge \big[‘\mathit{Req}\ ‘\mathit{Has\text{-}fur}\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\big] \wedge$$

$$\big[‘\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \big[‘= [[‘\mathit{Average}\ ‘\mathit{Skul\text{-}Size}]_{\mathrm{wt}}\,x]\ ’41.25\big]\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\big] \wedge$$

$$\big[‘\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \big[[‘\mathit{Ter\text{-}Marking}_{wt}\ x\ ‘\mathit{Clawing}] \vee [‘\mathit{Ter\text{-}Marking}_{wt}\ x\ ‘\mathit{Urinating}] \vee [‘\mathit{Ter\text{-}Marking}_{wt}\,x\ ‘\mathit{Droppings}]\big]\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\big] \wedge$$

$$\big[‘\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \big[‘= [‘\mathit{Pregnancy\text{-}Period}_{wt}\,x]\ ‘65\big]\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\big] \wedge$$

$$\big[‘\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \big[‘\leq [‘\mathit{Litter\text{-}Size}_{wt}\,x]\ ‘4\big]\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\big]$$

Source 5.

- The average body length of a wild cat is 47 cm or more.
- Wild cats mark their territory with claws, urination, droppings.
- The pregnancy period of a wild cat is 65 days.
- The size of the litter of wild cats is up to 4 kittens.

Exp (Source 5, ‘*Wild-cat*).

$$\big[‘\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \big[‘\geq [[‘\mathit{Averige}\ ‘\mathit{Body\text{-}Length}]_{\mathrm{wt}}\,x]\ ’47\big]\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\big] \wedge$$

$$\big[‘\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \big[[‘\mathit{Ter\text{-}Marking}_{wt}\,x\ ‘\mathit{Clawing}] \vee [‘\mathit{Ter\text{-}Marking}_{wt}\,x\ ‘\mathit{Urinating}] \vee [‘\mathit{Ter\text{-}Marking}_{wt}\,x\ ‘\mathit{Droppings}]\big]\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\big] \wedge$$

$$\big[‘\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \big[‘= [‘\mathit{Pregnancy\text{-}Period}_{wt}\,x]\ ‘65\big]\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\big] \wedge$$

$$\big[‘\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \big[‘\leq [‘\mathit{Litter\text{-}Size}_{wt}\,x]\ ‘4\big]\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\big]$$

Source 6.

- The body length of wild cats is 47 cm or more.
- Wild cat marks its territory with its claws, urination, droppings.
- Wild cats seek their mate by a loud meow.
- The size of the litter of wild cats is up to 4 kittens.

Exp (Source 6, ‘*Wild-cat*).

$$\Big[‘\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \Big[‘\geq [[‘\mathit{Averige}\ ‘\mathit{Body\text{-}Length}]_{\mathrm{wt}}x]\ ’47\Big]\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\Big] \wedge$$

$$\Big[‘\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \Big[[‘\mathit{Ter\text{-}Marking}_{wt}x\ ‘\mathit{Clawing}] \vee [‘\mathit{Ter\text{-}Marking}_{wt}x\ ‘\mathit{Urinating}] \vee [‘\mathit{Ter\text{-}Marking}_{wt}x\ ‘\mathit{Droppings}]\Big]\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\Big] \wedge$$

$$\Big[‘\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \Big[‘\mathit{Seek}_{wt}x\ ‘\mathit{Mate}\ [‘\mathit{Loud}\ ‘\mathit{Meow}]\Big]\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\Big] \wedge$$

$$\Big[‘\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \Big[‘\leq [‘\mathit{Litter\text{-}Size}_{wt}x]\ ‘4\Big]\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\Big]$$

Source 7.

- Wild cats are mammals.
- The weight of a wild cat is up to 11 kilograms.
- Wild cats usually live in mixed or deciduous forests.
- Wild cat marks its territory with its claws, urination, droppings.
- Wild cat looks for a mate with a loud meow.
- The pregnancy period of a wild cat is 65 days.
- Wild cat has fur.

Exp (Source 7, ‘*Wild-cat*).

$$\Big[‘\mathit{Req}\ ‘\mathit{Mammal}\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\Big] \wedge$$

$$\Big[‘\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \Big[‘\leq [‘\mathit{Weight}_{wt}x]\ ‘11\Big]\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\Big] \wedge$$

$$\Big[‘\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \Big[‘\mathit{Live\text{-}in}_{wt}\ x\ \lambda w\lambda t\ \lambda y\ [[[‘\mathit{Mixed}\ ‘\mathit{Forest}]_{wt}y] \vee [[‘\mathit{Decidious}\ ‘\mathit{Forest}]_{wt}y]]\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\Big] \wedge$$

$$\Big[‘\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \Big[[‘\mathit{Ter\text{-}Marking}_{wt}x\ ‘\mathit{Clawing}] \vee [‘\mathit{Ter\text{-}Marking}_{wt}x\ ‘\mathit{Urinating}] \vee [‘\mathit{Ter\text{-}Marking}_{wt}x\ ‘\mathit{Droppings}]\Big]\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\Big] \wedge$$

$$\Big[‘\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \Big[‘\mathit{Seek}_{wt}x\ ‘\mathit{Mate}\ [‘\mathit{Loud}\ ‘\mathit{Meow}]\Big]\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\Big] \wedge$$

$$\Big[‘\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \Big[‘{=}\ [‘\mathit{Pregnancy\text{-}Period}_{wt}x]\ ‘65\Big]\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\Big] \wedge$$

$$\Big[‘\mathit{Req}\ ‘\mathit{Has\text{-}fur}\ [‘\mathit{Wild}\ ‘\mathit{Cat}]\Big]$$

Source 8.

- The body length of wild cats is up to 80 cm.
- The size of the territory of a wild cat is greater than 50 ha.
- The size of the litter of wild cats is up to 4 kittens.

Exp (Source 8, ‘*Wild-cat*).

$$\left[{}'\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \left[{}'\leq [[{}'\mathit{Average}\ {}'\mathit{Body\text{-}Length}]_{wt}\,x]\ {}'80\right]\ [{}'\mathit{Wild}\ {}'\mathit{Cat}]\right] \wedge$$

$$\left[{}'\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \left[{}'\geq [{}'\mathit{Territory\text{-}Size}_{wt}\,x]\ {}'50\right]\ [{}'\mathit{Wild}\ {}'\mathit{Cat}]\right] \wedge$$

$$\left[{}'\mathit{Typ\text{-}p}\ \lambda w\lambda t\ \lambda x\ \left[{}'\leq [{}'\mathit{Litter\text{-}Size}_{wt}\,x]\ {}'4\right]\ [{}'\mathit{Wild}\ {}'\mathit{Cat}]\right]$$

Types:
$Wight, Body\text{-}Lenght, Height, Skull\text{-}Size, Territory\text{-}Size, Litter\text{-}Size,$
$In\text{-}Heat\text{-}Period, Pregnancy\text{-}Period/(\tau\iota)_{\tau\omega}$: attributes
$Average/((\tau\iota)_{\tau\omega}(\tau\iota)_{\tau\omega})$: attribute modifier
$Mammal, Cat, Has\text{-}Fur, Forests, Clawing, Urinating, Droppings,$
$Mate, Meow/(o\iota)_{\tau\omega}$: properties
$Wild, Loud, Mixed, Deciduous/((o\iota)_{\tau\omega}(o\iota)_{\tau\omega})$: property modifiers
$x \rightarrow \iota$
$Live\text{-}in, Ter\text{-}Marking/(o\iota(o\iota)_{\tau\omega})_{\tau\omega}$
$Seek/(o\iota(o\iota)_{\tau\omega}(o\iota)_{\tau\omega})_{\tau\omega}$
$Typ\text{-}p/(o(o\iota)_{\tau\omega}(o\iota)_{\tau\omega})$
$Req/(o(o\iota)_{\tau\omega}(o\iota)_{\tau\omega})$

Table 1 is the incident matrix computed from these explications.

Table 1. Incident matrix. Explications/properties

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
e_1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
e_2	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0
e_3	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
e_4	1	1	0	0	0	0	1	0	0	0	1	0	0	0	1	1	0
e_5	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	1	0
e_6	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	1	0
e_7	1	1	1	0	0	0	0	0	1	0	1	0	0	1	1	0	0
e_8	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0

1. $'Mammal$
2. $'Has$-fur
3. $\lambda w\ \lambda t\ \lambda x\ ['\leq [' Weight_{wt}\ x\ ]\ '11\ ]$
4. $\lambda w\ \lambda t\ \lambda x\ ['\geq [' Weight_{wt}\ x\ ]\ '1.2\ ]$
5. $\lambda w\ \lambda t\ \lambda x['\geq [['Average\ 'Body\text{-}Length]_{wt}\ x]\ \ '47]$
6. $\lambda w\ \lambda t\ \lambda x['\leq [['Average\ 'Body\text{-}Length]_{wt}\ x]\ \ '80]$
7. $\lambda w\ \lambda t\ \lambda x\ ['= [['Average\ 'Skul\text{-}Size\ ]_{wt}\ x]\ '41.25\ ]$
8. $\lambda w\ \lambda t\ \lambda x\ ['= [['Average\ 'Height\ ]_{wt}\ x]'37.6]$

9. $\lambda w\,\lambda t\,\lambda x\left['\mathit{Live\text{-}in}_{wt}\;x\;\left[\lambda w\,\lambda t\,\lambda y\,[[['\mathit{Mixed}\;\;'\mathit{Forrest}\,]_{wt}\;y]\;\vee\right.\right.$
$\left.\left.[['\mathit{Deciduous}\;\;'\mathit{Forrest}\,]_{wt}\;y]]\right]\right]$
10. $\lambda w\,\lambda t\,\lambda x\;['\geq['\mathit{Territory\text{-}Size}_{wt}\;x]\;'50\,]$
11. $\lambda w\,\lambda t\,\lambda x\left[['\mathit{Ter\text{-}Marking}_{wt}\;x\;\;'\mathit{Clawing}\,]\;\vee\right.$
$['\mathit{Ter\text{-}Marking}_{wt}\;x\;\;'\mathit{Urinating}\,]\;\vee$
$\left.['\mathit{Ter\text{-}Marking}_{wt}\;x\;'\mathit{Leaves\text{-}Droppings}\,]\right]$
12. $\lambda w\,\lambda t\,\lambda x\;['\leq['\mathit{In\text{-}Heat\text{-}Period}_{wt}\;x\,]\;'8]$
13. $\lambda w\,\lambda t\,\lambda x\;['\geq['\mathit{In\text{-}Heat\text{-}Period}_{wt}\;x\,]\;'2]$
14. $\lambda w\,\lambda t\,\lambda x\;\left['\mathit{Seek}_{wt}\;x\;'\mathit{Mate}\;['\mathit{Loud}\;'\mathit{Meow}\,]\right]$
15. $\lambda w\,\lambda t\,\lambda x\;['=['\mathit{Pregnancy\text{-}Period}_{wt}\;x]\;\;'65\,]$
16. $\lambda w\,\lambda t\,\lambda x\;['\leq['\mathit{Litter\text{-}Size}_{wt}\;x\,]\;'4]$
17. $\lambda w\,\lambda t\,\lambda x\;['\geq['\mathit{Litter\text{-}Size}_{wt}\;x\,]\;'3]$

Min-supp = 0.25
Min-conf = 0.66

Assume that the user has chosen the first explication as the basic one. Hence, the concepts corresponding to the columns 1-8 can occur only in the antecedents of the recommendation rules. The remaining concepts occur only in rule consequents.

Rules:
Confidence = 0.66; ***RS = {s4, s7}***
$\{'\mathit{Mammal},'\mathit{Has\text{-}fur}\}\;\Rightarrow_{e1}$
$\left\{\lambda w\,\lambda t\,\lambda x\;\left[\left['\mathit{Ter\text{-}Marking}_{wt}\;x\;\;'\mathit{Clawing}\,\right]\;\vee\;\left['\mathit{Ter\text{-}Marking}_{wt}\;x\;'\mathit{Urinating}\,\right]\right.\right.$
$\left.\left.\vee\left['\mathit{Ter\text{-}Marking}_{wt}\;x\;'\mathit{Leaves\text{-}Droppings}\,\right]\right]\right\}$

Confidence = 0.66; ***RS = {s4, s7}***
$\{'\mathit{Mammal},'\mathit{Has\text{-}fur}\}\;\Rightarrow_{e1}$
$\left\{\lambda w\,\lambda t\,\lambda x\;\left['=\left['\mathit{Pregnancy\text{-}Period}_{wt}\;x\right]'65\,\right],\lambda w\,\lambda t\,\lambda x\;\left[\left['\mathit{Ter\text{-}Marking}_{wt}\;x\;\;'\mathit{Clawing}\,\right]\right.\right.$
$\left.\left.\vee\;\left['\mathit{Ter\text{-}Marking}_{wt}\;x\;'\mathit{Urinating}\,\right]\;\vee\left['\mathit{Ter\text{-}Marking}_{wt}\;x\;'\mathit{Leaves\text{-}Droppings}\,\right]\right]\right\}$

Confidence = 0.66; ***RS = {s4, s7}***
$\{'\mathit{Mammal},'\mathit{Has}\text{-fur}\}\;\Rightarrow_{e1}$
$\{\lambda w\,\lambda t\,\lambda x\;['=['\mathit{Pregnancy\text{-}Period}_{wt}\;x]\;\;'65\,]\}$

Confidence = 0.75; ***RS = {s5, s6, s8}***
$\{\lambda w\,\lambda t\,\lambda x['\geq[['\mathit{Average}\;'\mathit{Body\text{-}Length}]_{wt}\;x]'47]\}\;\Rightarrow_{e1}$
$\{\lambda w\,\lambda t\,\lambda x\;['\leq['\mathit{Litter\text{-}Size}_{wt}\;x\,]\;'4]\}$

Based on the first explication e_1, the algorithm proposes textual resources as being relevant for the concept of wild cat. According to the first three rules, the algorithm proposes sources No. 4 and 7 because these documents contain information on mammals, those that have fur, on territory marking and pregnancy period. The last rule is a recommendation for the documents No. 5, 6 and 8; these sources contain information on average body length and litter size.

If the algorithm computed also weakly recommended documents (WRS) then it would not take into account the properties of being a mammal, having a fur and average body length, and thus it would recommend much more documents containing information on, for instance, territory marking.

6. Conclusion

In this paper we described the proposal of exploration of the data mining method of 'association rules' for the search of relevant textual documents. The goal is the selection of information sources that contain information relevant for dealing with or explaining or answering the initiative query on a simple concept *C*. The paper broadens our previous results on explication of simple concepts by means of molecular concepts extracted from textual documents by supervised machine learning methods. By applying these methods to textual resources, we obtain several explications of the simple input concept, which are further evaluated and processed. We introduced an algorithm that computes associations of the concepts occurring in these explications with other concepts from other resources. In this way the algorithm discovers hidden associations that might be relevant with respect to the query on the simple input concept; as a result, it recommends other textual resources that might be overlooked in the huge amount of input documents and thus ignored. Future research will concentrate on optimisation of this method, in particular on an effective generating of association rules from a large dataset obtained from a huge number of textual documents.

Concerning the entire project on natural language processing and question answering of which this system is a component, we will concentrate on improvement of the methods introduced here. In particular, molecular concepts that explicate a simple input concept and that are obtained from several textual resources should be checked for inconsistencies that contradict each other or yield paradoxes. Another promising idea seems to be checking the concepts of propositions for striking news that go against our common sense and intuitions. In this way, we can signalise fake news coming from unreliable Internet sources.

Acknowledgements. This research has been supported by the Grant Agency of the Czech Republic, project No. GA18-23891S, "Hyperintensional Reasoning over Natural Language Texts", and by The Ministry of Education,Youth and Sports of the Czech Republic from the National Programme of Sustainability (NPU II) project IT4Innovations Excellence in Science, LQ1602.

References

1. Agrawal, R., Imielinski, T., and Swami, A. N. (1993): Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD *International Conference on Management of Data*, pp. 207-216.
2. Carnap, R. (1947): *Meaning and necessity*. Chicago: Chicago University Press.

3. Duží, M. (2017): Property modifiers and intensional essentialism. *Computación y Sistemas,* vol. 21, No. 4, pp. 601–613. DOI: 10.13053/CyS-21-4-2811.
4. Duží, M., Fait, M., Menšík, M. (2019): Adjustment of goal-driven resolution for natural language processing in TIL. In *Recent Advances in Slavonic Natural Language Processing*, *RASLAN 2019*, Horák A., Rychlý P., Rambousek, A. (eds.), pp. 71-82.
5. Duží, M., Fait, M.: Integrating special rules rooted in natural language semantics into the system of natural deduction**.** In the proceedings of *ICAART 2020*, the *12th International Conference on Agents and Artificial Intelligence*, Ana Rocha, Luc Steels, Jaap van der Herik (eds.), vol.1, pp. 410-421, Malta, Valletta.
6. Duží, M., Jespersen, B., Materna, P. (2010): *Procedural Semantics for Hyperintensional Logic. Foundations and Applications of Transparent Intensional Logic*. Berlin: Springer.
7. "Globalisation: Threat or Opportunity?". International Monetary Fund, 12 April 2000. Retrieved 28 January 2020.
8. Hájek P., Havránek T., Chytil M.K. (1983): *Metoda GUHA - automatická tvorba hypotéz*. (In Czech. *GUHA method*; automatic creation of hypotheses). Academia Praha.
9. Medveď, M., Šulganová, T., Horák, A. (2017): Multilinguality Adaptations of Natural Language Logical Analyzer. In Proceedings of the Eleventh Workshop on *Recent Advances in Slavonic Natural Language Processing, RASLAN 2017*. Brno: Tribun EU, pp. 51-58.
10. Menšík, M., Duží, M., Albert, A., Patschka, V., Pajr, M. (2019): Seeking relevant information sources. In *Informatics'2019*, IEEE 15th International Scientific Conference on Informatics, Poprad, Slovakia, pp. 271-276.
11. Menšík, M., Duží, M., Albert, A., Patschka, V., Pajr, M. (2020): Machine learning using TIL. In *Frontiers in Artificial Intelligence and Applications, vol. 321*: *Information Modelling and Knowledge Bases XXXI*, A. Dahanayake, J. Huiskonen, Y. Kiyoki, B. Thalheim, H. Jaakkola, N. Yoshida (eds.), pp. 344-362, Amsterdam: IOS Press.
12. Menšík, M., Duží, M., Albert, A., Patschka, V., Pajr, M. (2020): Refining concepts by machine learning. *Computación y Sistemas*, Vol. 23, No. 3, pp. 943–958; doi: 10.13053/CyS-23-3-3242
13. Mitchell T. M. (1997): *Machine Learning*. New York: McGraw-Hill, 1997.
14. Moschovakis, Y. N. (1994): Sense and denotation as algorithm and value. In *Lecture Notes in Logic*, eds. J. Väänänen and J. Oikkonen, vol. 2, pp. 210-249. Berlin: Springer.
15. Poole D. L., Mackworth A. K. (2010): *Artificial Intelligence: Foundations of Computational Agents*. 2nd pub. Cambridge: Cambridge University Press.
16. Russell S. J., Norvig P.(2014): *Artificial intelligence: a modern approach*. 2nd ed. Harlow: Pearson Education, 2014. ISBN 978-1-29202-420-2.
17. Tichý, P. (1988): *The Foundations of Frege's Logic*. Berlin, New York: De Gruyter.
18. Winston P. H.(1992): *Artificial intelligence*. 3rd ed., Mass.: Addison-Wesley Pub. Co., 1992.