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

doi:10.3233/FAIA210487

Improvement of Searching for Appropriate Textual Information Sources Using Association Rules and FCA

Marek MENŠÍK [a], Adam ALBERT [a], Vojtěch PATSCHKA [a] Miroslav PAJR [b]
[a] *Department of Computer Science, FEI,*
VŠB - Technical University of Ostrava, 17. listopadu 15, 708 00 Ostrava,
Czech Republic
[b] *Institute of Computer Science,*
Silesian University in Opava, Berzučovo nám. 13, 746 01 Opava,
Czech Republic

Abstract. This paper deals with an optimization of methods for recommending relevant text sources. We summarize methods that are based on a theory of Association Rules and Formal Conceptual Analysis which are computationally demanding. Therefore we are applying the 'Iceberg Concepts', which significantly prune output data space and thus accelerate the whole process of the calculation. Association Rules and the Relevant Ordering, which is an FCA-based method, are applied on data obtained from explications of an atomic concept. Explications are procured from natural language sentences formalized into TIL constructions and processed by a machine learning algorithm. TIL constructions are utilized only as a specification language and they are described in numerous publications, so we do not deal with TIL in this paper.

Keywords. Association Rules, FCA, Iceberg Lattices, Relevant Ordering

1. Introduction

In case of studying certain problematic area, we need to acquire a list of appropriate papers we want to study to have the whole picture of a particular problem. Therefore in [1], [2] and [3], we introduced methods, where we utilize the methods of Association Rules and the Relevant Ordering based on the Formal Concept Analysis as a theoretical background for selecting the most relevant text-sources. Those methods are based on applying the theory of machine learning and concept explications (more in [4]). Because sentences in the natural language must be formalised into a formal language, we chose to avail of the strong system of Transparent Intensional Logic [5].

The main issue we need to deal with is the time complexity. Making the entire Concept Lattice is immensely time consuming so we were seeking for some time-optimization. Numerous approaches exist to the problem, so we chose to utilize *Iceberg Concepts*. The entire process is based on a horizontal space reduction where we cut a significant part of the concept lattice.

The paper is structured as follows. In chapter 2 we briefly introduce the problem of concept explication which is crucial for the next data processing. In chapter 3 we outline the theories applied in Association Rules, Formal Concept Analysis and Iceberg Lattices. The complete process of finding the set of recommended text sources is demonstrated by an example. For a clear comparison, we used the same example as in [2] and [3]. We point at some problems which might occur using our methods in combination with Iceberg Concepts. Chapter 5 concludes our paper.

2. Explication of an atomic concept

Since we deal with the natural-language processing, we use TIL as our background theory. TIL allows us to formalize salient semantic features of the natural language in a fine-grained way. For more details, see [5].

Atomic concepts are explicated by combination of TIL and machine learning. Explications provide understanding and additional useful information about atomic concepts. *Carnapian explication*[1] is the process of refinement of inaccurate or vague expression. The expression, to be refined, is called an *explicandum*; its refinement, obtained by the explication, is called an *explicatum*. For example, a simple expression such as a dog (explicandum) can be refined as "*Dog is a domesticated carnivore*" (explicatum). In terms of TIL, the explicandum is an atomic concept, i.e. an atomic closed construction. The explicatum is a molecular construction describing the explicandum. We also say that the molecular concept is an ontological definition of the object falling under the atomic concept.

For example:

$$'Dog \approx_{exp} \lambda w \lambda t \, \lambda x \, [['Domesticated \; 'Carnivore]_{wt} \, x]$$

$$\text{Types: } Domesticated/((o\iota)_{wt}(o\iota)_{wt}); Dog, Carnivore/(o\iota)_{wt}; x \to \iota$$

The algorithm of obtaining explications has been introduced in [4]. It exploits a symbolic method of supervised machine learning adjusted to the natural language processing. The input of the algorithm are sentences in natural language mentioning the expression to be explicated formalised as TIL constructions.

The algorithm, based on Patrick Winston's work [7], iteratively builds the explicatum using the constructions marked as positive or negative examples. With positive examples, we refine the explicatum by inserting new constituents into the molecular the construction or we generalize the explicatum so that can adequately define the explicandum. With negative examples, we specialize the explicatum by inserting new constituents in the negated way. By those constituents we differentiate the explicatum of our expression from similar expression's explicata.

[1] See [6].

3. Theoretical background

3.1. Association Rules

The method of Association Rules extraction has been introduced in [8]. Yet ten years earlier a similar method has been described in [9]. Basically, it is the process of looking for interesting relations among the large amount of data 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 data items occurring in a dataset that satisfy a predefined minimum support and confidence. The algorithm first extracts *k-frequent* item-sets, i.e. those item-sets whose occurrences exceed a predefined threshold k (minimal support). Then a *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 a more solid ground, here are the definitions. First, we need to define The *support* of a given set $\{i_1,\ldots,i_n\}$ of data items. It is the probability of an occurrence of the record with all these items in the dataset.

Definition 1 (*support, k-frequent item-set*). Let $I=\{i_1,\ldots,i_n\}$ be a set of data items and $D=\{T_1,\ldots,T_m\}$ a dataset of records such that each $T_i\subseteq I$. Then *support* of a set of items (item-set) $A\subseteq I$ in D is

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

The set A is *k-frequent item-set* iff $supp(A)\geq k$.

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

Definition 2 (*confidence of an association rule*). Let $I=\{i_1,\ldots,i_n\}$ be a set of data items and $D=T_1,\ldots,T_m$ a dataset of records such that each $T_i\subseteq I$. Then *association rule* is of the form $A\Rightarrow B$, where $A,B\subseteq I$ and $A\cap B=\emptyset$ and $A,B\neq\emptyset$. *Confidence* of the rule $A\Rightarrow B$ is

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

Definition 3 (*recommendation rule*). 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 $Prop(e_i)$ be the set of all constituents occurring in an explication $e_i\in E$. Then the *rule* $A\Rightarrow_e B$ is a *recommendation rule* for a given 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 recommendation rule for the given explication e.

Definition 4 (*recommended sources*). Let $A \Rightarrow_e B$ be a *recommendation rule* for an explication e. Let *exp(d,c)* be an explication of an input atomic 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

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

This 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.

3.2. Formal Concept Analysis

Formal Conceptual Analysis[2] (FCA) was introduced in 1980s by the group of researchers lead by Rudolf Wille and became a popular technique within the information retrieval field. FCA has been applied in many disciplines such as software engineering, machine learning, knowledge discovery and ontology construction. Informally, FCA studies how objects can be hierarchically grouped together with their mutual common attributes. The following definitions of *significant objects* and *relevant ordering* are originally presented in [3].

Definition 5 (*formal context*). Let B be a non empty finite set of objects, let M be non empty finite set of attributes and let I be a binary relation $I \subseteq G \times M$ called *incidence* that expresses which objects have which attributes. Then (G,M,I) is called *formal context*.

Definition 6 (*formal concept, extent, intent*). Let (G,M,I) be a formal context, then $\beta(G,M,I) = \{(O,A) | O \subseteq G, A \subseteq M, A^{\downarrow} = O, O^{\uparrow} = A\}$ is a set of all formal concepts of the context (G,M,I) where , $O^{\uparrow} = \{a | \forall o \in O, (o,a) \in I\}, A^{\downarrow} = \{o | \forall a \in A, (o,a) \in I\}$. $A^{\downarrow}$ is called *extent* of a formal concept (O,A) and $O^{\uparrow}$ is called *intent* of a formal concept (O,A).

Definition 7 *(significant objects)*. The set of *Significant objects* of an object e in $\beta(G,M,I)$ is a set $SO(e) = \bigcup_{i=1}^{n} O_i^e$, where O^e is an extent of a concept $(O,A) \neq (G,B)$, $e \in O$ and $B \subseteq M$. Hence, the set of significant objects of an object e is the union of all the extents which the object e is an element of.

[2]More in [10].

Definition 8 ***(relevant ordering)*.** Let $SO(e)$ be a set of all significant objects of an object e, let $\gamma(e)$ be a set of all concepts (O,A) where $e \in O$,i.e.: $\gamma(e) = \{(O^e,(O^e)^\uparrow)|(O^e,(O^e)^\uparrow) \neq (G,B), B \subseteq M, (O^e,(O^e)^\uparrow) \in \beta(G,M,I)\}$, then $\mathbf{a} \sqsubseteq \mathbf{b}$ is in a *relevant ordering*[3] iff

$$max(|(O^a)^\uparrow|) \leq max(|(O^b)^\uparrow|), a,b, \in SO(e), (O^a,(O^a)^\uparrow), (O^b,(O^b)^\uparrow) \in \gamma(e).$$

3.3. Iceberg Concept Lattices

Iceberg Concept Lattices [11] consist only of the top-most concepts of the concept lattice. Iceberg Concept Lattice is defined as follows:

Definition 9 (*iceberg concept lattice*). Let $(A,B) \in \beta(G,M,I)$ and let $supp(B) \geq min\text{-}supp$, then (A,B) is called *frequent concept*. The set of all frequent concepts of the context (G,M,I) is called the *Iceberg Concept Lattice* of the context (G,M,I)

According to the definition, the ICL represents the top part of the lattice as it is shown in Fig. 1.

4. Demonstration by an Example

As an example of recommending relevant information sources based on FCA, we use the same dataset we used in [2]. In our example, we used text sources dealing with the concept of *wild cat*. We obtained 8 explications of the concept from different textual sources $(s_1,...,s_8)$. Therefore each explication describes the concept of *being a wild cat* from the different point of view. To illustrate the basic idea without troubling the reader with too many technicalities, we present just one of those eight explications:

$e_1 = [\prime Typ\text{-}p\ \lambda w \lambda t\ \lambda x[[\prime\leq [\prime Weight_{wt}\ x]\ \prime 11] \wedge [\prime\geq [\prime Weight_{wt}\ x]\ \prime 1.2]][\prime Wild\ \prime Cat]] \wedge$
$[\prime Req\ \prime Mammal\ [\prime Wild\ \prime Cat]] \wedge [\prime Req\ \prime Has\text{-}fur\ [\prime Wild\ \prime Cat]] \wedge [\prime Typ\text{-}p\ \lambda w \lambda t\ \lambda x[[\prime\leq$
$[[\prime Average\ \prime Body\text{-}Length]\ x]\ \prime 80] \wedge [\prime\geq [[\prime Average\ \prime Body\text{-}Length]\ x]\ \prime 47]][\prime Wild\ \prime Cat]] \wedge$
$[\prime Typ\text{-}p\ \lambda w \lambda t\ \lambda x[[\prime= [[\prime Average\ \prime Skull\text{-}Size]\ x]\ \prime 41.25]][\prime Wild\ \prime Cat]] \wedge [\prime Typ\text{-}p\ \lambda w \lambda t\ \lambda x[[\prime=$
$[[\prime Average\ \prime Height]\ x]\ \prime 37,6]][\prime Wild\ \prime Cat]]$

Explication mentioned above was obtained from text source describing the wild cat from the biological point of view. It contained information such as classification of this specimen (being a mammal), body proportions and appearance of the wild cat.

After obtaining all explications, the user selects one of them which is the most relevant from his point of view. Let e_1 be the case. The entire process of recommendation starts after the explication selection.

From the explications mentioned above, we generate an incidence matrix written in Table 1.

Each row of the table represents one explication and each column represents particular property/attribute. Value 1 in a cell means that the respective explication contains the respective property, 0 otherwise.

The $e_1,...,e_8$ are identifiers of explications.

The columns' numbers in Table 1 represent the following attributes:

[3]Classical concept ordering is defined as: $(O,A) \sqsubseteq (O_1,A_1)$ iff $A \subseteq A_1$

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. $\lambda w\lambda t\lambda x['\geq [['Average\ 'Body\text{-}Length]\ x]\ '47]$
6. $\lambda w\lambda t\lambda x['\leq [['Average\ 'Body\text{-}Length]\ x]\ '80]$
7. $\lambda w\lambda t\lambda x['= [['Average\ 'Skull\text{-}Size]\ x]\ '41.25]$
8. $\lambda w\lambda t\lambda x['= [['Average\ 'Skull\text{-}Height]\ x]\ '37.6]$
9. $\lambda w\lambda t\lambda x['Live\text{-}in_{wt}\ [\lambda w\lambda t\lambda y[[['Mixed\ 'Forrest]_{wt}\ y]\ \vee\ [['Deciduous\ 'Forrest]_{wt}\ y]]]]$
10. $\lambda w\lambda t\lambda x['\geq ['Territory\text{-}Size_{wt}\ x]\ '50]$
11. $\lambda w\lambda t\lambda x[['Ter\text{-}Marking_{wt}\ x\ 'Clawing] \vee ['Ter\text{-}Marking_{wt}\ x\ 'Urinating] \vee$
    $['Ter\text{-}Marking_{wt}\ x\ 'Leaves\text{-}Droppings]]$
12. $\lambda w\lambda t\lambda x['\leq ['In\text{-}Heat\text{-}Periodwt\ x]\ '8]$
13. $\lambda w\lambda t\lambda x['\geq ['In\text{-}Heat\text{-}Periodwt\ x]\ '2]$
14. $\lambda w\lambda t\lambda x['Seek_{wt}\ x\ 'Mate\ ['Loud\ 'Meow]]$
15. $\lambda w\lambda t\lambda x['= ['Pregnancy\text{-}Periodwt\ x]\ '65]$
16. $\lambda w\lambda t\lambda x['\leq ['Litter\text{-}Size_{wt}\ x]\ '4]$
17. $\lambda w\lambda t\lambda x['\geq ['Litter\text{-}Size_{wt}\ x]\ '3]$

O/A	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
$\mathbf{e_1}$	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
$\mathbf{e_2}$	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0
$\mathbf{e_3}$	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
$\mathbf{e_4}$	1	1	0	0	0	0	1	0	0	0	1	0	0	0	1	1	0
$\mathbf{e_5}$	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	1	0
$\mathbf{e_6}$	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	1	0
$\mathbf{e_7}$	1	1	1	0	0	0	0	0	1	0	1	0	0	1	1	0	0
$\mathbf{e_8}$	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0

Table 1. Incident matrix

The *incident matrix* and the selected explication e_1 are are the common inputs for both the methods of recommendation, namely the one based on Association Rules and the one based on the Relevant Ordering

4.1. Recommendations based on Association Rules

Min-supp = 0.25 Min-conf = 0.66

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

Item-sets meeting the min-supp condition, i.e. *0.25-frequent item-sets*, are the following ones:

1. $\{1\}$
2. $\{1, 2\}$
3. $\{1, 2, 3\}$
4. $\{1, 2, 7\}$
5. $\{1, 2, 11\}$
6. $\{1, 2, 11, 15\}$
7. $\{1, 2, 15\}$
8. $\{1, 3\}$
9. $\{1, 7\}$
10. $\{1, 11\}$
11. $\{1, 11, 15\}$
12. $\{1, 15\}$

13. {2}
14. {2, 3}
15. {2, 7}
16. {2, 11}
17. {2, 11, 15}
18. {2, 15}
19. {3}
20. {5}
21. {5, 11}
22. {5, 11, 16}
23. {5, 16}
24. {7}
25. {9}
26. {9, 11}
27. {10}
28. {11}
29. {11, 14}
30. {11, 15}
31. {11, 15, 16}
32. {11, 16}
33. {14}
34. {14, 15}
35. {14, 16}
36. {15}
37. {15, 16}
38. {16}

Frequent item-sets which can be transformed into the rules where the antecedent contains only columns 1-8 and succedent 9-17 are these:
$\{1,11\}$, $\{1,11,15\}$, $\{1,15\}$, $\{1,2,11\}$,$\{1,2,11,15\}$,$\{1,2,15\}$,$\{2,11\}$,$\{2,11,15\}$, $\{2,15\}$,$\{5,16\}$

Final recommendation rules found according to our data are presented in table 2:

Rule	RS	Rule	RS
$\{1\} \Rightarrow_{e_1} \{11\}$	{s4}	$\{1,2\} \Rightarrow_{e_1} \{11,15\}$	{s4,s7}
$\{1\} \Rightarrow_{e_1} \{11,15\}$	{s4,s7}	$\{2\} \Rightarrow_{e_1} \{11\}$	{s4,s7}
$\{1\} \Rightarrow_{e_1} \{15\}$	{s4,s7}	$\{2\} \Rightarrow_{e_1} \{11,15\}$	{s4,s7}
$\{1,2\} \Rightarrow_{e_1} \{11\}$	{a4,s7}	$\{2\} \Rightarrow_{e_1} \{15\}$	{s4,s7}
$\{1,2\} \Rightarrow_{e_1} \{11,15\}$	{s4,s7}	$\{5\} \Rightarrow_{e_1} \{16\}$	{s5,s6, s8}

Table 2. Recommendation rules: min-supp = 0.25, min-conf = 0.6

Based on the first explication, the algorithm proposes other explications and thus also textual sources as relevant for the concept of wild cat. According to the rules, the algorithm recommends sources No. 4 and 7 because these documents contain information about territory marking and pregnancy period. The last rule is a recommendation of documents No. 5, 6 and 8; these sources contain information about litter size.

At this point, we can raise the min-supp up to 0.3 and compare the results. We can see that there are approximately 1/3 of all frequent item-sets[4] compare to the level set up to 0.25.

Min-supp = 0.3 Min-conf = 0.66

1. {1}
2. {1, 2}
3. {2}
4. {5}
5. {5, 16}
6. {11}
7. {11, 15}
8. {11, 16}
9. {14}
10. {15}
11. {15, 16}
12. {16}

The frequent item-set which can be transformed into the rule (antecedent contains only columns 1-8 and succedent 9-17) is the only one {5, 16}.

With increased min-supp value, we acquired only one association rule. Therefore the rule recommends documents No. 5, 6 and 8.

[4]It is clear that raising the min-supp number means that the final amount of frequent item-sets can not be higher.

Rule	RS
$\{5\} \Rightarrow_{e_1} \{16\}$	{s5,s6, s8}

Table 3. Recommendation rule min-supp = 0,3, min-conf = 0,6

In case of using Association Rules, min-supp adjustment is not always ideal optimization of the computation process. The best optimization for this method would be an optimization of generating of the frequent item-sets.

4.2. Relevant ordering based on FCA

From Table 1, by using FCA, we obtained the following concepts:

0. $(\{e_1,e_2,e_3,e_4,e_5,e_6,e_7,e_8\},\emptyset)$
1. $(\{e_1,e_4,e_7\},\{1,2\})$
2. $(\{e_1,e_4\},\{1,2,7\})$
3. $(\{e_1,e_5,e_6,e_8\},\{5\})$
4. $(\{e_1,e_7\},\{1,2,3\})$
5. $(\{e_1\},\{1,2,3,4,5,6,7,8\})$
6. $(\{e_2,e_4,e_5,e_6,e_7\},\{11\})$
7. $(\{e_2,e_7\},\{9,11\})$
8. $(\{e_2,e_8\},\{10\})$
9. $(\{e_2\},\{9,10,11\})$
10. $(\{e_3,e_4,e_5,e_6,e_8\},\{16\})$
11. $(\{e_3,e_4,e_5,e_7\},\{15\})$
12. $(\{e_3,e_4,e_5\},\{15,16\})$
13. $(\{e_3,e_6,e_7\},\{14\})$
14. $(\{e_3,e_6\},\{14,16\})$
15. $(\{e_3,e_7\},\{14,15\})$
16. $(\{e_3\},\{12,13,14,15,16,17\})$
17. $(\{e_4,e_5,e_6\},\{11,16\})$
18. $(\{e_4,e_5,e_7\},\{11,15\})$
19. $(\{e_4,e_5\},\{11,15,16\})$
20. $(\{e_4,e_7\},\{1,2,11,15\})$
21. $(\{e_4\},\{1,2,7,11,15,16\})$
22. $(\{e_5,e_6,e_8\},\{5,16\})$
23. $(\{e_5,e_6\},\{5,11,16\})$
24. $(\{e_5\},\{5,11,15,16\})$
25. $(\{e_6,e_7\},\{11,14\})$
26. $(\{e_6\},\{5,11,14,16\})$
27. $(\{e_7\},\{1,2,3,9,11,14,15\})$
28. $(\{e_8\},\{5,10,16\})$
29. $(\emptyset,\{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17\})$

Conceptual lattice of these formal concepts is visualised in Fig. 1. Concepts marked in SOC area contain only *significant objects*. The nodes with bright numbers represent particular explications. At this point, the reader does not have to care about the vertex colours.

Significant objects of the object (explication) e_1 is the following set:
$SO(e_1) = \{e_1,e_4,e_5,e_6,e_7,e_8\}$.

The set of all concepts containing explication e_1 as a common object is the following set:

$$\gamma(e_1) = \{(\{e_1\},\{1,2,3,4,5,6,7,8\}),(\{e_1,e_4\},\{1,2,7\}),(\{e_1,e_7\},\{1,2,3\}),(\{e_1,e_5,e_6,e_8\},\{5\}),(\{e_1,e_4,e_7\},\{1,2\}),\}$$

Formal concepts mentioned in the set $\gamma(e_1)$ are represented by numbers 1,2,3,4,5 in Fig. 1.

The *relevant ordering*[5] (defined in chapter 3.2) is represented by the following se-

[5]More details can be found in [3]

quence:

$$e_8(s_8) \sqsubseteq e_6(s_6) \sqsubseteq e_5(s_5) \sqsubseteq e_7(s_7) \sqsubseteq e_4(s_4) \sqsubseteq e_1(s_1)$$

4.3. Iceberg optimization

In this chapter, we deal with optimization of the above described methods. In general, there are thousands of vertexes or Association Rules in our graphs. It it not necessary to make the entire computation, because the majority of the computed data is irrelevant to user. Hence, it is plausible to reduce the space to some reasonable degree. To this end, we decided to apply Iceberg Concept Lattices. The whole process of finding the Iceberg Concept Lattice consists of two parts. The first one is finding *k-frequent item sets*. These are the sets of attributes which have the minimal support greater or equal to k. Those *k-frequent item-sets* are then used to find concepts and *recommended sources*.

As (Fig. 1) illustrates, there are numerous vertexes that are useless with respect to the selected explication e_1. Key part of the lattice is highlighted by SOC set. Concept No. 5 represents our explication and concepts No. 1,2,3,4 contain in their *extents* explications of the above mentioned recommended sources (e_4, e_7, e_5, e_6, e_8).

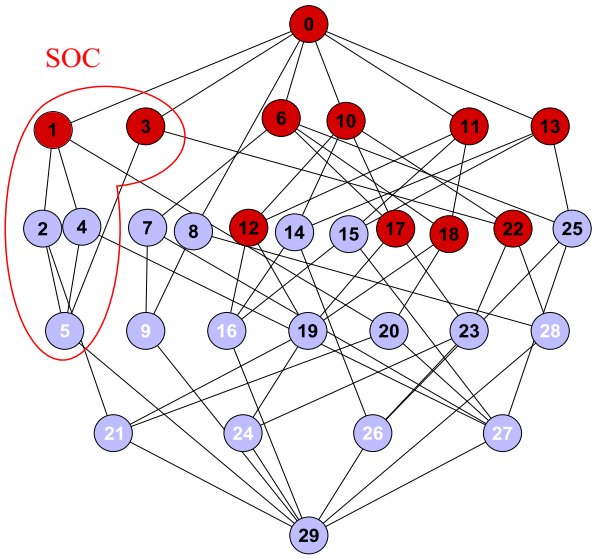

Figure 1. Iceberg lattice of formal concepts - dark vertexes

The entire optimization is based on rising of the *min-support* level. If *min-support* = 0 then *Iceberg lattice* is the same as the standard formal concept lattice. Raising to 0.25, there will by only 38 *frequent item-sets* and 21 concepts. But if we raise the *min-support* to 0.3, then the amount of final *frequent item-sets* will be significantly reduced. In our case, there would be just 12 frequent item-sets and 11 concepts left(highlighted by dark

vertexes in Fig. 1).

At this point, with the *Relevant ordering* algorithm, the *Significant objects* of the object (explication) e_1 is the following set: $SO(e_1) = \{e_1, e_4, e_5, e_6, e_7, e_8\}$. The set of all concepts containing our explication e_1 as a common object is the following set:

$$\gamma(e_1) = \{(\{e_1, e_4, e_7\}, \{1,2\}), (\{e_1, e_5, e_6, e_8\}, \{5\})\}$$

Formal concepts, mentioned in the set $\gamma(e_1)$, are represented by numbers 1, 3 in Fig. 1. As one can realize, the concept No. 5 $(\{e_1\}, \{1,2,3,4,5,6,7,8\})$ is not in $\gamma(e_1)$. It is not that important, because the particular explication e_1 was selected by user. Thus the user is aware of existing e_1 and the relevance would be the highest one. At this point we can show the final *relevant ordering*:

Exp.	Intent	DF	RT
e_1	{1,2,3,4,5,6,7,8}	{}	$\{s_1\}$
e_4	{1,2}	{3,4,5,6,7,8}	$\{s_4\}$
e_7	{1,2}	{3,4,5,6,7,8}	$\{s_7\}$
e_5	{5}	{1,2,3,4,6,7,8}	$\{s_5\}$
e_6	{5}	{1,2,3,4,6,7,8}	$\{s_6\}$
e_8	{5}	{1,2,3,4,6,7,8}	$\{s_8\}$

Table 4. Final text sources' ordering. Min-supp = 0.3

As we can see above, the result will be same as the result using the entire conceptual lattice:

$$e_8(s_8) \sqsubseteq e_6(s_6) \sqsubseteq e_5(s_5) \sqsubseteq e_7(s_7) \sqsubseteq e_4(s_4) \sqsubseteq e_1(s_1)$$

As stated earlier, this method of Iceberg Latices is not generally applicable as the optimization method on Association Rules, because raising the min-support can lead to the loss of an important information. But if we use FCA, the method significantly reduces the data space that is necessary to deal with.

5. Conclusion

In this paper we have summarized two previously proposed methods for text source recommendation. As mentioned in [3], [2], we needed to focus on making recommendation more reliable and computationally more effective. Therefore in this paper, we applied the method exploiting Iceberg Lattices. Recommending method exploiting the Association Rules [2], seeks on the basis of the specified min-supp and min-conf. Using Iceberg Lattices is not suitable as an improvement of this method. Raising min-supp leads to a loss of a large amount of relevant data and information. The best optimization for this method would be an optimization of the frequent item-sets generation.

Recommending method based on FCA [3] utilizes the *Relevant Ordering* of documents. Using Iceberg Lattices as an improvement of this method has proved to be effective, as there is only vertical reduction of data space, thus no loss of information oc-

curred. The loss of information would occur only with a substantial increase in min-supp, thus causing no results to be obtained at all. By an example, we demonstrated how our methods work, and both the methods were implemented in our SW application.

The only obstacle, shared by both the methods, is that they require *explications* of individual atomic concepts from given text sources.It is challenging to automate this process. We are working on the improvement of the transfer of sentences in the natural language into the language of the TIL constructions.

Acknowledgements

This research has been supported by Grant of SGS No. SP2021/87, VSB - Technical University of Ostrava, Czech Republic, “Application of Formal Methods in Knowledge Modelling and Software Engineering IV” and also this work was supported by ESF project ’Zvýšení kvality vzdělávání na Slezské univerzitě v Opavě ve vazbě na potřeby Moravskoslezského kraje’ CZ.02.2.69/0.0/0.0/18_05 8/0010238.

References

[1] Menšík, M., Duží, M., Albert, A., Patschka, V., Pajr, M. (2019): Seeking relevant information sources. In Informatics’2019, *IEEE 15th International Scientific Conference on In-formatics*, Poprad, Slovakia, pp. 271-276.

[2] Albert, A., Duží, M., Menšík, M., Pajr, M., Patschka, V. (2021): Search for Appropriate Textual Information Sources. In *Frontiers in Artificial Intelligence and Applications*, vol. 333: Information Modelling and Knowledge Bases XXXII, B. Thalheim, M. Tropmann-Frick, H. Jaakkola,N. Yoshida, Y. Kiyoki (eds.), pp. 227-246, Amsterdam: IOS Press, doi: 10.3233/FAIA200832

[3] Menšík, M., Albert, A., Patschka, V. (2020): Using FCA for Seeking Relevant Information Source. In *RASLAN 2020*, Brno: Tribun EU, 2020, 144 p. ISBN 978-80-263-1600-8, ISSN 2336-4289.

[4] Menšík, M., Duží, M., Albert, A., Patschka, V., Pajr, M. (2019): Refining concepts by machine learning.*Computación y Sistemas*, Vol. 23, No. 3, 2019, pp. 943–958, doi: 10.13053/CyS-23-3-3242

[5] Duží, M., Jespersen, B., Materna, P. (2010): Procedural Semantics for Hyperintensional Logic. *Foundations and Applications of Transparent Intensional Logic*. Berlin: Springer.

[6] Carnap, Rudolf (1964): *Meaning and Necessity: A Study in Semantics and Modal Logic*. Chicago: University of Chicago Press.

[7] Winston P. H.(1992): *Artificial intelligence*. 3rd ed., Mass.: Addison-Wesley Pub. Co., 1992.

[8] 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.

[9] Hájek P., Havránek T., Chytil M. K. (1983): *Metoda GUHA - automatická tvorba hypotéz.* (In Czech. GUHA method; automatic creation of hypotheses). Academia Praha.

[10] Ganter, B., Wille, R. (1999): *Formal Concept Analysis: Mathematical Foundations*. 1st ed., Berlin: Springer. ISBN 978-3-540-62771-5.

[11] Taouil, R., Bastide, Y., Lakhal, L. (2001): *Conceptual Clustering with Iceberg Concept Lattices*.