Journal of Siberian Federal University. Mathematics & Physics 2017, 10(3), 372–384

УДК 519.7

On Algebraic Approach of R. Wille and B. Ganter in the Investigation of Texts

Valentina V. Bykova*
Institute of Mathematics and Computer Science,
Siberian Federal University,
Svobodny, 79, Krasnoyarsk, 660041,
Russia

Choduraa M. Mongush†
Tuvan State University
Lenina, 36, Kyzyl, Tuva, 667000,
Institute of Mathematics and Computer Science,
Siberian Federal University,
Svobodny, 79, Krasnoyarsk, 660041,
Russia

Received 17.03.2017, received in revised form 20.04.2017, accepted 20.05.2017

The statement of the problem of a binary classification by precedents using formal concept lattices is given, in which the initial data are two binary contexts. It is specified that this problem is intractable due to the high computational complexity of discovery process of the formal concept and constructing for them of the lattices. The decomposition reception, which allows reducing the computational complexity of this process is proposed and theoretically justified. The reduction of computational complexity is achieved by separation of every initial context on polynomial number of boxes (subcontexts), followed by a search of the formal concepts in each selected box. The results of computational experiments are presented and they confirm the effectiveness of the proposed of reception of the reducing computational complexity.

Keywords: text mining, classification, Galois connection, formal concepts lattice, decomposition context.
DOI: 10.17516/1997-1397-2017-10-3-372-384.

Introduction

Nowadays, the corpuses of natural languages of the peoples of the Russian Federation are created actively for the preservation of the national literary heritage and conducting of scientific research on the studying of languages. In the frames of corpuses a lot of problems of analysis of natural language texts, arising in the philological and linguistic investigations are solved by the application of mathematical methods. One of them is the problem of the classification by precedents, which allows establishing a genre of the work, of the author of the work, and of the space-time period of writing work. An algebraic approach is based on the formal concept analysis can be used for solution of these problems. Within this approach texts of works are modelled by contexts and are presented by binary matrices, reflecting the ratio of the presence or absence of attributes specific to the studied corpus of works.

*bykvalen@mail.ru
†mongushchod91@yandex.ru

Formal Concept Analysis (FCA) is an applied branch of the algebraic theory of lattices, in frames of which a formalization of the terms "concept" and "hierarchy of concepts" [1, 2] is possible. The basic ideas of FCA were formulated in the works of R. Wille and B. Ganter [3, 4] and are developed in the works of S. O. Kuznetsov, D. I. Ignatov, S. I. Gurov [5–8]. In FCA the term "concept" is determined using Galois connections and is a pair of the sets (extension, intention). The main advantage of this definition is full correspondence with the traditional interpretation of the term "concept", which is used in the humanities [9, 10].

The algebraic approach of R. Wille and B. Ganter has found wide application in various areas of machine learning [11–16]. Traditionally the classification of natural language texts is carried out based on quantitative proximity measures of the considered texts [17]. An approach of R. Wille and B. Ganter allows to classify natural language texts in qualitatively level instead of quantitatively level, i.e. through the presence or absence of the characteristic attributes. Moreover, the identified and related into the lattice the formal concepts are of special value since they define the conceptual model of the investigated field [18].

In this paper the necessary information about Galois connection and lattices of closed sets, the formal statement of a binary classification problem in terms of the FCA are given. The main result is presented, it is the decomposition reception which allows to reduce the computational complexity of the process of solving this problem. The correctness of the proposed decomposition reception is proved.

1. The main provisions of formal concept analysis

Firstly, let's give the definitions and notions of the FCA for facilitate the comprehension of the content of this paper. They were taken from [1, 3] and are described in common notation.

A binary relation $\sqsubseteq$ on set of P is the ratio of (non-strict) partial order if it holds for all $x, y, z \in P$ the following properties:

- *reflexive*: $x \sqsubseteq x$,
- *antisymmetry*: if $x \sqsubseteq y$ and $y \sqsubseteq x$, then $x = y$,
- *transitivity*: if $x \sqsubseteq y$ and $y \sqsubseteq z$, then $x \sqsubseteq z$.

The set P with certain on it by the ratio partial order $\sqsubseteq$ is called the partially ordered set (or further poset) and is denoted by $(P, \sqsubseteq)$. An upper bound of set $X \subseteq P$ in a poset $(P, \sqsubseteq)$ is an element $a \in P$, such that $x \sqsubseteq a$ for all $x \in X$. The least (or smallest) upper bound of the set X, denoted by the $\sup X$, is such its upper bound of a, that $a \sqsubseteq b$ for any upper bound b of this set. Dually, defines the notion $\inf X$, i.e. the least (or greatest) lower bound of the set $X \subseteq P$. A lattice is called a poset L in which any two elements x and y have a greatest lower bound (or meet denoted by $x \sqcap y$), and a least upper bound (or join denoted by $x \sqcup y$). A lattice L is complete when each of its subset X has a least upper bound and a greatest lower bound in L.

Let two non-empty finite set G and M objects and attributes are defined for certain a subject domain, respectively (from the German words *Gegenstande* is the object, *Merkmale* is an attribute). Let us assume that all objects in the G and attributes in M are different. Let there be given an incidence relation $I \subseteq G \times M$ between the sets G and M. A triple $K = (G, M, I)$ is called a formal context for considered subject domain. Existence in I of a pair $(g,\ m)$, $g \in G$ and $m \in M$, means that the object g has attribute m and vice versa, the attribute m is inherent object g. Further for brevity we shall omit the word "formal" and a triple $K = (G, M, I)$ will

simply be called context. The objects and attributes of the context $K = (G, M, I)$ will be called its elements.

If the sets G and M linearly ordered (for example, lexicographical)

$$G = \left\{g_1, g_2, \ldots, g_{|G|}\right\},$$

$$M = \left\{m_1, m_2, \ldots, m_{|M|}\right\},$$

then any context $K = (G, M, I)$ can be uniquely (up to a "material" nature of the objects and attributes) set by an incidence matrix $T_K = \|t_{ij}\|$, where

$$t_{ij} = \begin{cases} 1, & \text{if } (g_i, m_j) \in I, \\ 0, & \text{if } (g_i, m_j) \notin I, \end{cases}$$

$(i = 1, 2, \ldots, |G|;\ j = 1, 2, \ldots, |M|)$. The matrix T_K is called object-attribute matrix of the context $K = (G, M, I)$. Further, let us assume that the context $K = (G, M, I)$ is uniquely represented by the matrix T_K.

Let us choose in $K = (G, M, I)$ two arbitrary elements: the object $g \in G$ and an attribute $m \in M$. Let's define the two mappings φ and ψ for them as follows:

$$\varphi(g) = \{m \in M \mid (g, m) \in I\},$$

$$\psi(m) = \{g \in G \mid (g, m) \in I\},$$

where $\varphi(g)$ is a set of attributes inherent to the object $g \in G$, and $\psi(m)$ is a set of objects that have attribute $m \in M$. Mappings φ and ψ are easily generalized to a set of objects $A \subseteq G$ and a set of attributes $B \subseteq M$ as follows:

$$\varphi(A) = \bigcap_{g \in A} \varphi(g) = \{m \in M \mid \forall\, g \in A\ (g, m) \in I\},$$

$$\psi(B) = \bigcap_{m \in B} \psi(m) = \{g \in G \mid \forall\, m \in B\ (g, m) \in I\}.$$

Thus, $\varphi(A)$ is a set of attributes that are common to all objects of A, and $\psi(B)$ is a set of objects which have all attributes from B. Mappings φ and ψ are defined such that if $A_1, A_2 \subseteq G$ and $B_1, B_2 \subseteq M$, then

$$\varphi(A_1 \cup A_2) = \varphi(A_1) \cap \varphi(A_2),$$

$$\psi(B_1 \cup B_2) = \psi(B_1) \cap \psi(B_2).$$

It is reasonable to put that $\varphi(\varnothing) = M$ and $\psi(\varnothing) = G$: empty set of objects has all the attributes from the M, every object the considered a context $K = (G, M, I)$ possesses empty set of attributes.

According to the traditions of the FCA for mappings φ and ψ is used a single designation $(\cdot)'$, and listed above formulas for the $\varphi(A), \psi(B)$ are recorded as follows:

$$A' = \bigcap_{g \in A} g' = \{m \in M \mid \forall\, g \in A\ (g, m) \in I\}, \tag{1}$$

$$B' = \bigcap_{m \in B} m' = \{g \in G \mid \forall\, m \in B\ (g, m) \in I\}. \tag{2}$$

If $g \in G$ and $m \in M$, then the designations g' and m' are usually serve an abbreviated form of the record of the sets $\varphi(g) = \{g\}'$ and $\psi(m) = \{m\}'$ respectively.

Mappings $'$ are peculiar a number of properties arising from their definition and quite realistic and postulated in the analysis of the provisions of the data: expansion (reduction) of set of features reduces (increases) the number of objects having these attributes. Formally, these properties can be expressed as the following statements.

Proposition 1. *For each context $K = (G, M, I)$ and any subsets $B_1, B_2 \subseteq M$ the next properties are correct:*

- *antimonotony: if $B_1 \subseteq B_2$, then $(B_2)' \subseteq (B_1)'$;*
- *extensiveness: $B_1 \subseteq (B_1)''$, where $(B_1)'' = ((B_1)')' \subseteq M$.*

The set $(B_1)'' = \varphi(\psi(B_1))$ can be interpreted as a set of attributes that always appear in objects of the context of $K = (G, M, I)$, together with attributes from B_1.

Proposition 2. *For each context $K = (G, M, I)$ and any subsets $A_1, A_2 \subseteq G$ the next properties are correct:*

- *antimonotony: if $A_1 \subseteq A_2$, then $(A_2)' \subseteq (A_1)'$;*
- *extensiveness: $A_1 \subseteq (A_1)''$, where $(A_1)'' = ((A_1)')' \subseteq G$.*

The set $(A_1)'' = \psi(\varphi(A_1))$ can be interpreted as a class of similar objects, i.e. of the objects that are sure to have all the attributes inherent to the objects A_1. Moreover, this set is the greatest by inclusion within the context of $K = (G, M, I)$.

Let's remark that according to propositions 1 and 2 mappings φ and ψ are constituted a pair of Galois connections between sets 2^G and 2^M, partially ordered by set-theoretic inclusion [1, 3]. It is known that for the Galois connections φ and ψ following equalities are fair:

$$\varphi(\psi(\varphi(A))) = \varphi(A),\ \psi(\varphi(\psi(B))) = \psi(B)$$

or the same thing in common notations

$$\big((A')'\big)' = (A'')' = A',\ ((B')')' = (B'')' = B'.$$

The double application of the mapping $'$ defines a closure operator to 2^M in the algebraic sense. It is peculiar

- *reflexive*: for any $B \subseteq M$ always $B \subseteq B''$;
- *monotony*: if $B_1 \subseteq B_2 \subseteq M$, then $(B_1)'' \subseteq (B_2)'' \subseteq M$;
- *idempotency*: for any $B \subseteq M$ always $(B'')'' = B''$.

These properties are followed from propositions 1 and 2. Similarly, it can define the closure operator on 2^G. It is obvious that $M = M''$ and $G = G''$. From reflexive of the operator closure and antimonotony of mappings $'$ is implied the following proposition.

Proposition 3. *For any context $K = (G, M, I)$ and any $A \subseteq G$ and $B \subseteq M$, the inclusion of $A \subseteq B'$ is true if and only if $B \subseteq A'$.*

The set of attributes $B \subseteq M$, for which $B = B''$, is called the closed relatively of the operator $''$ in the context $K = (G, M, I)$. A decision has also to talk that the set B'' is the

closure for $B \subseteq M$ in a given context. From idempotency operator closure is followed that for any $B \subseteq M$ and every context $K = (G, M, I)$ the set B'' always is closed.

In view (1) and (2) the closure for $B \subseteq M$ relative to the given context $K = (G, M, I)$ can directly calculate from the formula:

$$B'' = (B')' = \varphi(\psi(B)) = \begin{cases} \left(\bigcap\limits_{m \in B} m'\right)' = \bigcap\limits_{g \in B'} g', & \text{if } B' \neq \varnothing, \\ M, & \text{if } B' = \varnothing. \end{cases} \tag{3}$$

In the FCA a pair of sets (A, B), $A \subseteq G$, $B \subseteq M$, such that $A' = B$ and $B' = A$, are called a formal concept of the context $K = (G, M, I)$. A set A is called an extent, B is called an intent of formal context K [3, 5]. Based on this definition, a pair of sets (A, B) is the formal concept in the context $K = (G, M, I)$ if and only if $A = A''$ and $B = B''$, i.e. when A, B are the closed sets relatively to the operator $''$ in $K = (G, M, I)$. If the context $K = (G, M, I)$ is represented by a matrix $T_K = \|t_{ij}\|$, then formal concept is corresponded the maximal its sub-matrix filled by units. The rows of this sub-matrix are corresponded to the elements from A, and the columns are corresponded to the elements from B.

Let FC_K is the set of all formal concepts of the context $K = (G, M, I)$. On FC_K the relation of partial order $\sqsubseteq$ through the set-theoretic inclusion we introduce as follows:

$$(A_1, B_1) \sqsubseteq (A_2, B_2), \text{ if } A_1 \subseteq A_2 \text{ (or } B_2 \subseteq B_1), \tag{4}$$

where $A_1, A_2 \subseteq G$ and $B_1, B_2 \subseteq M$. Let us note that in the statement (4) is sufficient to indicate only one of the two inclusions $A_1 \subseteq A_2$, $B_2 \subseteq B_1$, since by antimonotony of mappings $'$ from one of them always is followed other. According by (4), if $(A_1, B_1) \sqsubseteq (A_2, B_2)$ then formal concept $Y = (A_2, B_2)$ can be considered more general than the concept $X = (A_1, B_1)$.

We define the operations of intersection $\sqcap$ and union $\sqcup$ on FC_K through the same name set-theoretic operations $\cap$ and $\cup$ as follows:

$$(A_1, B_1) \sqcap (A_2, B_2) = (A_1 \cap A_2, (A_1 \cap A_2)'), \tag{5}$$

$$(A_1, B_1) \sqcup (A_2, B_2) = ((B_1 \cap B_2)', B_1 \cap B_2). \tag{6}$$

Then poset $(FC_K, \sqsubseteq)$ forms a lattice $L_K = (FC_K, \sqcap, \sqcup)$. Operations $\sqcap$ and $\sqcup$, defined by (5) and (6), are satisfied all the necessary laws for lattices associative, commutative, idempotency and absorption [1]. This lattice is called concept lattice of the context $K = (G, M, I)$. It is known that a lattice L_K is complete lattice [3]. The zero element of lattice L_K is a formal concept $\big(M', M\big)$, containing all the attributes of the context $K = (G, M, I)$. The unit of lattice L_K is a formal concept $\big(G, G'\big)$, in which the extent is the set of objects considered context.

2. Statement of the problem

Traditionally, the problem of binary classification by precedents is formulated in the following way [6, 17]. Let be given a finite set of objects G, divided into two classes G^+ and G^-, such that $G^+ \cap G^- = \varnothing$, $G^+ \cup G^- = G$. This division is determined by using some learning sample and the target binary attribute z. The elements of sets G^+ and G^- are called positive and negative precedents (learning step), respectively. All objects of G are described by a finite set of attributes M. This description is given by (0, 1)-matrix T_K, encoding the presence or absence

of attributes. Let be given an object $x \notin G$, described as through the set of attributes M. It is required to find the decision rule (or a rule classification), which on the basis of the matrix T_K for an object x determines the class to which it can be referred. The decision rule should lead to rejection of the classification when the question of ownership of the object x to a particular class is still open.

We formalize the problem of binary classification by precedent in FCA terms. Let us represented the classes G^+ and G^- of objects by two contexts in relation to the target binary attribute z: the positive context $K^+ = (G^+, M, I^+)$, the negative context $K^- = (G^-, M, I^-)$. Let $M_x \subseteq M$ be attributive description of the object $x \notin G$. Then the solution of the problem of a binary classification is reduced to finding a decision rule defining the class to which can be referred an object x.

There are various algorithms of the binary classification on based FCA. They are including: algorithms RULEARNER, GALOIS, GRAND, CITRIC, based on the use of all concepts lattice [12, 13], algorithms CLAN and CLUB, LEGAL, using some subset of the concepts lattice [14], and algorithms which are based on hypotheses [6, 15]. A visual representation of the results in the form of lattices is the advantage of these algorithms. The high computational complexity, is mainly determined by the size of the used lattice for formal concepts is the main disadvantage of these algorithms. In this paper, to solve the problem of a binary classification on precedents is used the algorithm which is based on hypotheses [15]. The choice of this algorithm is explained as follows. This algorithm instead of one lattice L_K of all formal concepts for investigated domain is working with two lattices L_{K^+} and L_{K^-} smaller size, constructed for a positive context and negative context, respectively. Lattices L_{K^+} and L_{K^-} allow to identify hypotheses and based on their perform the classification.

A hypothesis is called a set attributes which presents in the description of objects one class and absents in the description of objects of another class. Hypothesis are retrieved from lattices of formal concepts L_{K^+} and L_{K^-}. A content B^+ of a formal concept $\left(A^+, B^+\right) \in L_{K^+}$ is called a positive hypothesis if doesn't exist the formal concept $\left(A^-, B^-\right) \in L_{K^-}$ such that $B^+ \subseteq B^-$. Otherwise B^+ is called a false positive generalization. The negative hypothesis and the false negative hypothesis are defined the similarly: a content B^- of the formal concept $(A^-, B^-) \in L_{K^-}$ is called a negative hypothesis if doesn't exist the formal concept $(A^+, B^+) \in L_{K^+}$ such that $B^- \subseteq B^+$, otherwise B^- is considered the false negative hypothesis.

The decision rule of a binary classification for the object x can be formulated as follows [15]:

- the object x belongs to the class of G^+, if the set M_x includes at least one positive hypothesis and doesn't include any negative hypotheses. Otherwise, the object x belongs to the class of G^-;

- the rejection of classification occurs if M_x doesn't contain as the subsets both the positive and the negative hypotheses and if M_x contains both the positive and the negative hypotheses.

Usually, the quality of classification can be assessed according to criteria completeness and accuracy, and also according the special test sets [17]. Applied to the problems of analysis of natural language texts that can be solved within philological and linguistic researches, the correctness of constructed formal concept lattices and of performed on their basis of classification can be estimated by experts (philologists and linguists). This is explained, primarily, by that the in FCA the classification of text is performed instead of quantitative level is made on the

qualitative level, i.e. through the presence or absence of the characteristic attributes in the investigated texts.

3. The process of solving the problem

A solution of binary classification problem by precedents includes the following stages: the preprocessing of the contexts; the finding of the positive and negative formal concept; the building of the lattices L_{K^+} and L_{K^-}; the identifying of the hypotheses; the application of the classification rules for the object $x \notin G$.

At the first stage is performed the preprocessing the initial contexts $K^+ = (G^+,\ M,\ I^+)$ and $K^- = (G^-,\ M,\ I^-)$ in order to decrease their size. The reducing may involve as a set of objects, and a set of attributes of initial context. The preprocessing is performed so that the number and composition of the formal concepts in lattices L_{K^+} and L_{K^-} have not changed. In the second stage the positive and the negative formal concepts in the initial contexts, which have been preprocessed, are identifying. The simplest way of implement these actions is the enumeration of all the various subsets of the set attributes (their number is generally much less than the number of objects) with the computation of closure for each of them. With an algorithmic point of view for a finding of closures instead of the formula (3), more convenient is used the formula (7) which is presented in the following proposition.

Proposition 4. *If $B' \neq \varnothing$, then the closure B'' for $B \subseteq M$ relative to the context $K = (G,\ M,\ I)$ coincides with the intersection of all sets of attributes possessed by the objects the considered of context, and that contain all the attributes from B:*

$$B'' = \bigcap_{g \in G} \left\{ g' \mid B \subseteq g' \right\}. \tag{7}$$

Proof. Let $S = \left\{ g \in G \mid B \subseteq g' \right\}$ be a set of objects, possessing all the attributes from B, and may be by some other features and $S' = \bigcap\limits_{g \in S} g'$ is a set of attributes common to all objects in S. According to (3) at $B' \neq \varnothing$, we have

$$B'' = (B')' = \left(\bigcap_{m \in B} m' \right)' = \bigcap_{g \in B'} g',$$

where B' is the set of objects for which common are all attributes from B. It is required to prove that $S' = B''$. Indeed, from the definition of the set S implies that $B' \subseteq S$ and $B \subseteq S'$. According to proposition 3, the inclusion $B \subseteq S'$ is true if and only if $S \subseteq B'$. This means, $S = B'$ and therefore, $S' = B''$. Note that when $B' = \varnothing$ invariably $B'' = M$. □

Application of the formula (7) allows finding the closure B'' for a given set of attributes $B \subseteq M$ per one view context $K = (G,\ M,\ I)$. Note that according to (7) the extension of the context $K = (G,\ M,\ I)$ by adding new objects to G does not alter the closure B'', calculated relatively to the old context but may expand composition of objects that possess all of the attributes from B''. This means that at such transformation of context the formal concepts have been found earlier can be changed only in regard to an increase their extents. In addition, new formal concepts may experience.

At the third stage the positive and negative concepts are ordered according to (4) and are built the lattices L_{K^+} and L_{K^-} using the formulas (5) and (6).

At the fourth and fifth stages are identified the hypotheses (positive, negative and false generalizations) by checking of the relations of the inclusion of the intents of the corresponding of formal concepts. Then, in accordance with the given above by decision rule of the classification is accept decided to refer an object x to the positive or to the negative class, or specify that the classification is not possible (to state the rejection of classification).

The stages 2 and 3 of described above process of solving the problem of classification (finding the formal concept and construction of lattices) have a high computational complexity. It is known that the problem of generation all formal concepts of the context $K = (G, M, I)$ and the problem of construction of formal concept lattices are NP-hard. The rationale of this fact presented in [7]. The high computational complexity due to the fact, that the number of formal concept can be exponential from size of the context. For example, this is the case for contexts of the form $K = (G, G, \neq)$. Therefore, the efficiency of algorithms generation of formal concepts is accepted to estimate as a function of the output length, i.e. the number of formal concepts. According to this, the time required to identify all formal concepts of the context $K = (G, M, I)$ in the worst case is $O(|L_K| \cdot |G|^2 \cdot |M|)$. A decrease values of $|G|$ and $|M|$ at the stage of preprocessing allows in some cases reduce the computation time of all formal concepts of the context $K = (G, M, I)$.

4. The decomposition of context

In practice, it is reduced the computational complexity of stages 2, 3 of the process of solving the problem of binary classification by precedents can also by applying of the decomposition reception: the separation of the initial context on the polynomial number of boxes, followed by search of the formal concepts in each of the selected boxes.

Let us introduced the concept of box. An object concept of context $K = (G,\ M,\ I)$ is called formal concept of the form (g'', g'), where $g \in G$, and attribute concept is the formal concept of the form (m', m''), where $m \in M$. Thus, each object in G corresponds to the separate object concept, and each attribute of M corresponds some an attribute concept. For the context $K = (G,\ M,\ I)$ the number of object concepts is equal to $|G|$, and the number of attribute concepts is equal $|M|$. Note that the object concept (g'', g') has the largest by capacity the intent among the other of the formal concepts that have in the extent of the object g, and attribute concept (m', m'') has the largest by size the extent among the other of the concepts that have in the intent of the attribute m. This follows from the antimonotony of Galois connection.

Let us denoted by $O_K = \{(g'', g') \mid \forall g \in G\} \subseteq FC_K$ the set of all object concepts and by $S_K = \{(m', m'') \mid \forall m \in M\} \subseteq FC_K$ the set of all attribute concepts of the context $K = (G, M, I)$. Note that the set of O_K and S_K may have a non-empty intersection. Let's choose a pair of formal concepts, the first of which is an object and the second is an attribute: $(g'', g') \in O_K$ and $(m', m'') \in S_K$. If for this pair is true relation $(g'', g') \sqsubseteq (m', m'')$, or the same performed the following conditions

$$g'' \subseteq m' \ \text{ and } \ m'' \subseteq g', \tag{8}$$

then (m', g') is called box of the context $K = (G,\ M,\ I)$, which is constituted by the elements $g \in G$ and $m \in M$ this context. It is obvious that among boxes are possible the duplicates, i.e. the boxes with equal sets of objects and attributes. However, the number of different boxes for given context $K = (G, M, I)$ always not exceeds by $|G| \cdot |M|$.

Let's say that a formal concept $(A, B) \in FC_K$ is imbedded into the box (m', g') of context $K = (G,\ M,\ I)$, if $A \subseteq m'$, $B \subseteq g'$. According to (8) any box (m', g') is not empty, since

into him always is embedded by at least two formal concepts $(g'', g') \in O_K, (m', m'') \in S_K$, $(g'', g') \sqsubseteq (m', m'')$, if they are different, and one if they match.

Let us consider a certain box (m', g') of context $K = (G, M, I)$, formed by the elements $g \in G$ and $m \in M$. Obviously, that (m', g') determines a certain submatrix of the matrix T_K and forms subcontext $C = (G, M, I_C)$ of context $K = (G, M, I)$, where $I_C \subseteq I$. Herewith, $(x, y) \in I_C$ if and only if $x \in m'$ and $y \in g'$. It is remarkable that the matrix representing an incidence relation I_C, always has the rows filled by units, corresponding to objects g'', and the columns filled by units, corresponding to the attributes m''. This follows from the definition of the box (m', g'). Let $|m'| \cdot |g'|$ is the size of box (m', g'), and n is the number of the elements of the matrix is representing I_C which are equal to unit. The quantity

$$\sigma(m', g') = \frac{n}{|m'| \cdot |g'|}$$

is called the density of box (m', g'). For density of the box is true

$$0 < \sigma(m', g') \leqslant 1.$$

If $\sigma(m', g') = 1$, then box (m', g') contains exactly one formal concept (m', g') of context $K = (G, M, I)$. If $\sigma(m', g') < 1$, then the box contains several of formal concepts of that context. The correspondence between the boxes and the formal concepts of context $K = (G, M, I)$ establishes the following proposition.

Proposition 5. *For any context $K = (G, M, I)$ and any pair of sets (A, B), where $\varnothing \neq A \subseteq G$, $\varnothing \neq B \subseteq M$, are fair the following statements:*

1. *If (A, B) is the formal concept of context $K = (G, M, I)$, then always in this context there is a box (m', g'), formed by the elements $g \in G$ and $m \in M$, at that perhaps not the only one in which this formal concept is embedded;*

2. *If (X, Y) is the formal concept subcontext $C = (G, M, I_C)$, corresponding to a certain box (m', g') of context $K = (G, M, I)$, then it is also a formal concept of context $K = (G, M, I)$.*

Proof. Let us prove the first statement. Let (A, B) be an arbitrary formal concept, which is different from $(G, \varnothing)$ and $(\varnothing, M)$. By definition, for him are true the equalities:

$$(A, B) = (B', A') = (A'', B'')\,. \tag{9}$$

Let us consider the some object $g \in A$. Let us find g' for him, i.e. the set of attributes it possesses and corresponding to it object concept (g'', g'). Since $\{g\} \subseteq A$, then by virtue (9), the antimonotony of the mapping $'$ and the monotone of the closure operators are fair the relationships of the embedding

$$A' \subseteq g', \quad g'' \subseteq A''. \tag{10}$$

Similarly, for any attribute $m \in B$ we have the attribute concept (m', m'') and the relationship of the embedding

$$B' \subseteq m', \quad m'' \subseteq B''. \tag{11}$$

From (9)–(11) is follows directly the justice of the conditions (8): $g'' \subseteq m'$ and $m'' \subseteq g'$. Consequently, a pair (g'', g') and (m', m'') forms a box (m', g'). Moreover, also

$$A = B' \subseteq m', \quad B = A' \subseteq g'. \tag{12}$$

This means that the formal concept (A, B) is enclosed in a box (m', g'). It is obvious that if choose another object of A and other attribute of B, then we obtain the same box or perhaps another box containing the formal concept (A, B). Note that in (9)–(12), the mapping $'$ and operator $''$ are calculated relative context $K = (G, M, I)$.

If in the context $K = (G, M, I)$ have the formal concepts of $(G, \varnothing)$ and $(\varnothing, M)$, then impossible to build for them the attribute and object concepts, respectively. Therefore, they are not embedded in any of the boxes of context $K = (G, M, I)$. Hence, they must be added in constructing of the lattice L_K.

Let us now prove the second statement of proposition 5. Let (A, B) be the formal concept of subcontext $C = (G, M, I_C)$, corresponding to box (m', g'). Obviously that the given a formal concept is embedded in the (m', g'). Therefore, the inclusions are true for him: $A \subseteq m', B \subseteq g'$. Note that here the mappings $'$ are defined relative the context $K = (G, M, I)$. Further, in order to distinguish, relative what of the context are found these mappings, will be indicated the name of context in the lower the index, and write, for example, so: m'_K, g'_K. According to the condition of the second statement

$$(A, B) = (B'_C, A'_C) = (A''_C, B''_C), \tag{13}$$

$$A = A''_C \subseteq m'_K, \quad B = B'' \subseteq g'_K. \tag{14}$$

We assume the contrary. Let in the initial context $K = (G,\ M,\ I)$ have a formal concept $(X, Y) = (Y'_K, X'_K) = (X''_K, Y''_K)$ such that

$$A \subset X \ \text{ and } \ B \subset Y, \tag{15}$$

which is not a formal concept of box (m', g'). Then by the antimonotonicity of Galois connections and relations (13)–(14), the condition (15) contradicts the definition of boxing. After all, based on the definition of box (m'_K, g'_K) the m' is the largest extent among the formal concepts of context $K = (G, M, I)$, having the attribute m in the intent. And set of g' determines the greatest intent by capacity among the other the formal concepts of context $K = (G, M, I)$, having the object g in the extent. Hence, (A, B) is a formal concept of context $K = (G, M, I)$. $\square$

The single formation of boxes for a given context $K = (G, M, I)$ includes the following steps: the finding the set O_K of object concepts; the finding the set S_K of attribute concepts; the checking of the conditions (8) for each pair of formal concept from O_K, S_K and the formation of the boxes. The number of such pairs, the checks and the received boxes by not more than $|I| = |G| \cdot |M|$. For the construction of all the object and attribute concepts is necessary $O(|G| \cdot |I|)$ and $O(|M| \cdot |I|)$ the time, respectively. In general, the time of formation of boxes is polynomial and constitutes $O(|I| \cdot (|G| + |M|))$. In the worst case, only one box can be found, which coincides with the initial context, and then the decomposition of context on the boxes is not effective. Such a situation is possible, for example, for the context is completely filled by units. However, the real contexts are decomposed into a reasonable number of boxes, usually.

Note that the process of decomposition of the initial context on the boxes can be arranged in stages. Because each the identified box whose density is strictly less than 1, may be re-divided into boxes. However, if this process is to continue until all boxes are degenerate in the formal concepts, it may result in an exponential number of boxes, and hence also to an exponential time of their construction. For obtaining of the polynomial number of boxes follows limit oneself by constant number of iterations. Let's note that the concept of box is used by us is similar to such concepts as bicluster and co-cluster, which are used for grouping of objects in the field of gene expression [8].

5. The results of computational experiments

Algorithms of solutions problem of the binary classification and proposed above receptions of finding of the formal concepts have a program realization on object-oriented language Delphi 2010. For check of the effectiveness of the proposed of the decomposition reception were conducted the computing experiments on contexts reflecting the belonging of Tuvan folklore texts to the heroic epic genre. Contexts with the number of objects 15, 18, 20 and the number of attributes 15 were studied. For each of these contexts was carried finding the set of all formal concepts FC_K without separation and with the single separation of the initial context on boxes. The results of computational experiments are given in Tab. 1, where $|G|$ is the number of objects of the initial context $K = (G, M, I)$, $|FC_K|$ is the number of found formal concepts, N is the number of identified boxes, t is the time run of the program.

Table 1. The experimental results

The finding all formal concepts	$\|G\|$	$\|FC_K\|$	N	t, ms
Without separation on boxes	15	36	–	480
With separation on boxes		36	12	66
Without separation on boxes	18	73	–	12480
With separation on boxes		73	23	120
Without separation on boxes	20	98	–	30519
With separation on boxes		98	40	150

The computational experiments are showed that the number and composition of the obtained formal concepts coincide completely in both cases (without separation on boxes, with separation on boxes). However, the using of the boxes gives a considerable gain in time.

Conclusion

In this paper is proposed and theoretically justified the decomposition reception which allows in practice significantly reduce the computation time of formal concepts. The reduction is achieved by dividing of the initial context on polynomial number of boxes, followed by a search of the formal concepts in each selected box. The conducted the computing experiments confirm the effectiveness of this reception. The developed software tools that implement the decision of the binary classification problem based on FCA, are included in the National corpus of the Tuvan language.

The presented in the work the results can be applied for solving problems arising in the philological and linguistic investigations by classification and identification of natural language texts. Is further is assumed the creation and justification of the preprocessing procedures of initial context and the research of structure of the boxes, in order to increase the efficiency of the developed algorithms and programs.

This work was supported by the Russian Humanitarian Science Foundation, grant 16-34-1-01033.

References

[1] G.Birkhoff, Lattice Theory, AMS, Providence, 1967.

[2] P.Cimiano, A.Hotho, S.Staab, Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis, *Journal of Artificial Intelligence Research*, **24**(2005), 305–339.

[3] B.Ganter, R.Wille, Formal Concept Analyses: mathematical foundations, Springer Science and Business Media, 2012.

[4] B.Ganter, Two basic algorithms in concept analysis, Springer, 2010.

[5] S.O.Kuznetsov, Mathematical aspects of concept analysis, *Journal of Mathematical Sciences*, **80**(1996), no. 2, 1654–1698.

[6] S.I.Gurov, Boolean Algebras, Ordered Sets, Lattices: definitions, properties, examples, Moscow, KRASAND, 2012 (in Russian).

[7] S.O.Kuznetsov, Automatic learning based on formal concept analysis, *Automatizatsiya i Distantsionnoe Upravlenie*, **10**(2001), 3–27 (in Russian).

[8] D.I.Ignatov, S.O.Kuznetsov, J.Poelmans, Concept-based Biclustering for Internet Advertisement, Proceedings of the 12th International Conference on Data Mining Workshops, IEEE Computer Society, 2012, 123–130.

[9] F.A.Brokgauz, I.A.Efron, Philosophical Dictionary of logic, psychology, ethics, aesthetics and the history of philosophy, St.Peterburg, 1911 (in Russian).

[10] E.K.Voyshvillo, Understood as a form of thinking: logical-epistemological analysis, *International Journal of General Systems*, 1989 (in Russian).

[11] R.Belohlavek, B. De Baets, J.Outrata, V.Vychodil, Inducing decision trees via concept lattices, AMS, Providence, **38**(2009), no. 4, 455–467.

[12] M.Sahami, Learning classification Rules Using Lattices, *Proc ECML, Heraclion, Crete, Greece*, (1995), 343–346..

[13] C.Caprineto, G.Romano, GALOIS An order-theoretic approach to conceptual clustering, In proceedings of ICML93, Amherst, USA, (1993), 33–40.

[14] Z.Xie, W.Hsu, Z.Liu, M.Lee, Concept Lattice based Composite Classifiers for high Predictability, *Artificial Intelligence, Wollongong, Australia*, **139**(2002), 253–267.

[15] N.Meddouri, M.Meddouri, Classification Methods based on Formal Concept Analysis, *CLA*, (2008), 9–16.

[16] A.Neznanov , D.Ilvovsky , S.Kuznetsov, A New FCA-based System for Data Analysis and Knowledge Discovery, Contributions to the 11th International Conference on Formal Concept Analysis, Dresden, Germany, 2013, 31–44.

[17] A.A.Barseghyan, M.S.Kupriyanov, V.V.Stepanenko, I.I.Kholod, Data Analysis Technology: Data Mining, Visual Mining, Text Mining, OLAP, Piter, St.Peterburg, 2008 (in Russian).

[18] D.V.Vlasov, The methods of forming the theoretical concepts, *Zh. Buryat. Gos. Univ.*, (2009), no. 6, 37–41 (in Russian).

Об алгебраическом подходе Р. Вилле и Б. Гантера в исследовании текстов

Валентина В. Быкова
Институт математики и фундаментальной информатики
Сибирский федеральный университет
Свободный, 79, Красноярск, 660041
Россия

Чодураа М. Монгуш
Тувинский государственный университет
Ленина, 36, Кызыл, Республика Тыва, 667000
Институт математики и фундаментальной информатики
Сибирский федеральный университет
Свободный, 79, Красноярск, 660041
Россия

Приведена постановка задачи бинарной классификации по прецедентам с использованием решеток формальных понятий, в которой исходными данными выступают два бинарных контекста. Отмечено, что данная задача труднорешаема за счет высокой вычислительной сложности процесса выявления формальных понятий и построения для них решеток. Предложен и теоретически обоснован декомпозиционный прием, позволяющий снизить вычислительную сложность этого процесса. Снижение вычислительной сложности достигается за счет разделения всякого исходного контекста на полиномиальное число боксов (подконтекстов) с последующим поиском формальных понятий в каждом выделенном боксе. Представлены результаты вычислительных экспериментов, подтверждающие эффективность предложенного приема снижения сложности вычислений.

Ключевые слова: анализ естественно-языковых текстов, классификация, соответствие Галуа, решетка формальных понятий, декомпозиция контекста.