ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ В ДИСКРЕТНОЙ МАТЕМАТИКЕ

УДК 519.7

ДЕКОМПОЗИЦИОННЫЙ ПОДХОД К ИССЛЕДОВАНИЮ ФОРМАЛЬНЫХ КОНТЕКСТОВ

В. В. Быкова*, Ч. М. Монгуш**

*Сибирский федеральный университет, г. Красноярск, Россия

**Тувинский государственный университет, г. Кызыл, Республика Тыва, Россия

Исследуется #P-полная задача нахождения всех формальных понятий заданного контекста и предлагается декомпозиционный метод её решения. В качестве частей разложения предлагается использовать фрагменты исходного контекста, названные боксами. Доказано, что разделение контекста на боксы «безопасно» относительно формальных понятий: при декомпозиции ни одно формальное понятие не теряется и не появляются новые формальные понятия. Доказано, что число боксов, возникающих на каждой итерации разложения, равно числу единичных элементов 0,1-матрицы, представляющей исходный формальный контекст. Предлагается уменьшать число боксов на каждой отдельной итерации процесса декомпозиции с помощью построения взаимно непересекающихся цепей боксов. Приводятся результаты вычислительных экспериментов, свидетельствующие о существенном повышении производительности алгоритмов нахождения всех формальных понятий при применении предлагаемого декомпозиционного метода.

Ключевые слова: *анализ формальных понятий, декомпозиция формального контекста.*

DOI 10.17223/20710410/44/9

DECOMPOSITIONAL APPROACH TO RESEARCH OF FORMAL CONTEXTS

V. V. Bykova*, Ch. M. Mongush**

*Siberian Federal University, Krasnoyarsk, Russia

**Tuvan State University, Kyzyl, Tuva, Russia

E-mail: bykvalen@mail.ru, mongushchod91@yandex.ru

The problem of finding all formal concepts of a given formal context is investigated. The problem arises when data mining is presented in the form of a binary object-attribute matrix, i.e. a matrix the rows of which correspond to objects, and the columns correspond to features that take a value from the two-element set $\{0, 1\}$. Here the value 1 of the element of a matrix is interpreted as the presence of corresponding attribute to the object, and 0 is as its absence. Such a representation of a set of data allows to be used the algebraic approach of R. Wille and B. Ganter, known in the literature as Formal Concept Analysis. Within of this approach the

initial object-attribute matrix is called the formal context, and any of its maximal full submatrix is a formal concept. In the problem of finding all formal concepts, it is required to find the set of all formal concepts for a given formal context. This problem belongs to combinatorial enumeration problems and is #P-complete. The high computational complexity of the problem is due to the fact that in the general case the number of formal concepts exponentially depends on the size of the initial formal context. Currently many algorithms have been developed to solve the problem, among them NextClosure, Close-by-One, Norri. The execution time of these algorithms in the worst case exponentially depends on the dimension of the initial context, and therefore they are unsuitable for practical analysis of contexts of large dimension. We propose a decomposition method for solving the problem under consideration. In this method, some fragments of the initial context defined in a certain constructive way are called boxes. We prove that the division of context into boxes is "safety" relatively of the formal concepts. This means that the formal concepts are not lost and new formal concepts do not arise at decomposition. We prove that the number of boxes arising at each iteration of the decomposition is equal to the number of unit elements of the 0,1-matrix representing the initial formal context. We show how a partial order relation can be defined on a set of boxes. We also show that the number of boxes at each separate iteration of the decomposition process can be reduced by using constructing mutually disjoint chains of boxes. The results of computational experiments are given, indicating that the application of the proposed decomposition method allows significantly to increase the performance of algorithms for finding all formal concepts of a given context.

Keywords: *formal concept analysis, formal context decomposition.*

Введение

Во многих задачах интеллектуального анализа данных изучаемая предметная область описывается с помощью объектно-признаковой таблицы, в которой каждый столбец соответствует некоторому признаку, а каждая строка определяет признаковое описание отдельного объекта. Подобное представление множества данных позволяет при их обработке применять математический аппарат общей алгебры и теории графов, например, объектно-признаковые таблицы можно исследовать с помощью методов анализа формальных понятий. Анализ формальных понятий (АФП, англ. Formal Concept Analysis — FCA) как прикладное направление теории решеток Г. Биркгофа возникло с появлением работ Б. Гантера и Р. Вилле [1 – 3]. В АФП объектно-признаковая таблица представляется формальным контекстом, отражающим наличие или отсутствие признаков, характерных для изучаемого множества объектов, и моделируется 0,1-матрицей. Каждое формальное понятие определяется парой замкнутых множеств, интерпретируемых как объём и содержание этого понятия. В матричной форме формальному понятию соответствует некоторая максимально полная подматрица 0,1-матрицы, представляющей формальный контекст.

С помощью методов АФП решаются различные прикладные задачи, связанные с классификацией и кластеризацией данных, выявлением зависимостей между данными и семантическим анализом естественно-языковых текстов [4 – 7]. В них формальные понятия трактуются как перекрёстные ассоциации, кластеры или бикластеры. В рамках АФП решение указанных задач сводится к нахождению всех формальных понятий исходного формального контекста с последующим связыванием их в решётку. Полу-

ченная решётка служит концептуальной моделью исследуемой предметной области и основой для решения прикладных задач.

При всей привлекательности методов АФП их практическое применение ограничивается высокой трудоёмкостью процесса извлечения множества формальных понятий из исходного контекста. Задача нахождения всех формальных понятий формального контекста детально изучена в [2 – 9]. Известно, что данная задача относится к классу #P-полных задач, поскольку число формальных понятий может экспоненциально зависеть от размера исходного контекста [4]. В настоящее время актуальны исследования по повышению производительности алгоритмов установления всех формальных понятий за счёт параллельных вычислений и применения декомпозиционного подхода [3, 5]. Основная цель этих исследований — сделать более доступными методы АФП для анализа больших данных.

В данной работе рассматривается задача нахождения всех формальных понятий заданного контекста. Предлагается декомпозиционный метод её решения, в котором частями разложения выступают фрагменты исходного контекста, названные боксами. Такая декомпозиция позволяет разлагать контекст без потери формальных понятий и тем самым снижать время выполнения алгоритмов решения рассматриваемой задачи. Исследуется структура боксов с целью оценки числа фрагментов разложения, получаемых на каждой итерации декомпозиции. Устанавливаются правила остановки итерационного процесса разложения. Приводятся результаты вычислительных экспериментов, подтверждающих результативность предложенного метода декомпозиции.

1. Основные положения и обозначения

Рассмотрим основные положения и типовые обозначения АФП [2, 3].

Пусть определены два непустых конечных множества: множество объектов G (нем. Gegenstände) и множество признаков или свойств M (нем. Merkmale). Пусть также задано непустое отношение инцидентности $I \subseteq G \times M$. Данное отношение содержит информацию о выполнимости свойств из M на объектах из G, т. е. $(g, m) \in I$ означает, что объект g обладает признаком m, и наоборот — признак m присущ объекту g. Тройку $K = (G, M, I)$ принято называть формальным контекстом.

Далее будем полагать, что множества G и M линейно упорядочены (например, лексикографически). В этом случае формальный контекст $K = (G, M, I)$ однозначно задаётся 0,1-матрицей $T = (t_{ij})$: $t_{ij} = 0$ при $(g_i, m_j) \notin I$ и $t_{ij} = 1$ при $(g_i, m_j) \in I$ $(i = 1, 2, \ldots, |G|;\ j = 1, 2, \ldots, |M|)$.

Выберем в $K = (G, M, I)$ два произвольных элемента $g \in G$, $m \in M$ и определим для них отображения $(\cdot)'$ следующим образом:

$$g' = \{m \in M : (g, m) \in I\}, \quad m' = \{g \in G : (g, m) \in I\}. \tag{1}$$

Согласно этому определению, множество g' устанавливает набор признаков, присущих объекту g, а множество m' задает семейство объектов, обладающих признаком m. Отображения $(\cdot)'$ легко обобщаются на множества $A \subseteq G$ и $B \subseteq M$:

$$A' = \bigcap_{g \in A} g', \quad B' = \bigcap_{m \in B} m'.$$

Для всякого формального контекста $K = (G, M, I)$ и любых подмножеств $B_1, B_2 \subseteq M$ верны следующие свойства:

— *антимонотонность*: если $B_1 \subseteq B_2$, то $(B_2)' \subseteq (B_1)'$;

— *экстенсивность*: $B_1 \subseteq (B_1)''$, где $(B_1)'' = ((B_1)')' \subseteq M$.

Аналогичные свойства справедливы для подмножеств множества G. Известно, что отображения $(\cdot)'$ являются соответствиями Галуа и для них верны следующие равенства [1]:

$$\big((A')'\big)' = (A'')' = A', \quad ((B')')' = (B'')' = B'. \tag{2}$$

Двойное применение $(\cdot)'$ определяет оператор замыкания $(\cdot)''$ на 2^M в алгебраическом смысле. Этому оператору присущи следующие свойства:

— *рефлексивность*: для любого $B \subseteq M$ всегда $B \subseteq B''$;
— *монотонность*: если $B_1 \subseteq B_2 \subseteq M$, то $(B_1)'' \subseteq (B_2)'' \subseteq M$;
— *идемпотентность*: для любого $B \subseteq M$ всегда $(B'')'' = B''$.

Множество $(B)''$ можно трактовать как набор признаков, которые неизменно появляются в объектах формального контекста $K = (G, M, I)$ вместе с признаками из B, причём это множество является наибольшим по включению в пределах этого формального контекста. Если $B = B''$, то B называется замкнутым множеством относительно оператора $(\cdot)''$.

Пара множеств (A, B), $A \subseteq G$, $B \subseteq M$, таких, что $A' = B$ и $B' = A$, называется формальным понятием формального контекста $K = (G, M, I)$ с объёмом A и содержанием B. Далее в ряде случаев определение «формальный» перед словами «контекст» или «понятие» будет опускаться.

Из (2) и определения оператора $(\cdot)''$ вытекает справедливость следующего высказывания: пара множеств (A, B) является формальным понятием тогда и только тогда, когда $A = A''$ и $B = B''$. Очевидно также, что всякое формальное понятие уникально в заданном контексте, т. е. отличается от других формальных понятий объёмом и/или содержанием. Если формальный контекст представлен 0,1-матрицей T, то при $A \neq \varnothing$ и $B \neq \varnothing$ формальному понятию (A, B) отвечает максимальная полная подматрица матрицы T. Строки этой подматрицы соответствуют элементам из A, а столбцы — элементам из B. Здесь под полной подматрицей понимается подматрица, все элементы которой равны 1; полная подматрица является максимальной, если она не содержится в других полных подматрицах.

Обозначим через FC множество всех формальных понятий формального контекста $K = (G, M, I)$. Пусть $(A_1, B_1), (A_2, B_2) \in FC$. Множество FC частично упорядочено отношением

$$(A_1, B_1) \sqsubseteq (A_2, B_2)$$

тогда и только тогда, когда $A_1 \subseteq A_2$. Отметим, что последнее эквивалентно условию $B_2 \subseteq B_1$. Каждое формальное понятие $(A, B) \in FC$ определяет для исследуемой предметной области совокупность однородных объектов A со своим специфичным набором признаков B. Если в контексте $K = (G, M, I)$ нет признаков, которые присущи всем объектам из G, то множество FC содержит формальное понятие $(G, \varnothing)$. Если в контексте нет объектов, обладающих всеми признаками из M, то $(\varnothing, M) \in FC$. Если имеют место оба случая одновременно, то $(G, \varnothing) \in FC$ и $(\varnothing, M) \in FC$. Эти формальные понятия называются тривиальными.

По определению формального контекста $K = (G, M, I)$ отношение I не пустое. Следовательно, отвечающая формальному контексту матрица T всегда ненулевая, а соответствующее ему множество FC не является пустым.

Определим на FC операции пересечения $\sqcap$ и объединения $\sqcup$ через одноимённые теоретико-множественные операции $\cap$ и $\cup$ следующим образом:

$$(A_1, B_1) \sqcap (A_2, B_2) = (A_1 \cap A_2, (A_1 \cap A_2)'),$$
$$(A_1, B_1) \sqcup (A_2, B_2) = ((B_1 \cap B_2)', B_1 \cap B_2).$$

Тогда упорядоченное множество $(FC, \sqsubseteq)$ образует решётку $L = (FC, \sqcap, \sqcup)$, которая называется решёткой формальных понятий контекста $K = (G, M, I)$. Нулем решётки $L = (FC, \sqcap, \sqcup)$ является формальное понятие (M', M), содержащее все признаки контекста $K = (G, M, I)$, а единицей — формальное понятие (G, G'), в котором объём — множество всех объектов рассматриваемого контекста.

2. О задаче нахождения всех формальных понятий и родственных с ней задачах

В задаче нахождения всех формальных понятий требуется найти множество FC для заданного контекста $K = (G, M, I)$. Данная задача относится к комбинаторным перечислительным задачам и является #P-полной [4]. Высокая вычислительная сложность задачи обусловлена тем, что в общем случае число формальных понятий экспоненциально зависит от размера исходного контекста. Например, это имеет место для контекста вида $K = (G, G, \neq)$. Такому контексту соответствует 0,1-матрица, в которой все элементы равны единице, за исключением диагональных элементов. Легко убедиться, что такой контекст содержит ровно $2^{|G|}$ формальных понятий.

На сегодняшний день для определения множества FC и построения решётки $L = (FC, \sqcap, \sqcup)$ разработано много алгоритмов, в их числе NextClosure, Close-by-One, Norris [3 – 5]. Время их выполнения алгоритмов в худшем случае составляет $O(|FC| \cdot |G|^2 \cdot |M|)$. Поскольку величина $|FC|$ может экспоненциально зависеть от $|G|$ и $|M|$, то время выполнения также может быть экспоненциальным. Повысить производительность алгоритмов определения множества FC и построения решётки $L = (FC, \sqcap, \sqcup)$ можно за счёт параллельных вычислений и применения декомпозиционного подхода [3, 5].

Важно отметить, что задача нахождения всех формальных понятий контекста $K = (G, M, I)$ эквивалентна задаче определения всех максимальных полных подматриц 0,1-матрицы T, отвечающей этому контексту. Существуют и другие родственные с ней задачи, например задачи, связанные с нахождением биклик в заданном двудольном графе. В самом деле, бинарную матрицу T можно рассматривать в качестве матрицы смежности двудольного графа, две доли которого соответствуют множествам строк и столбцов матрицы T. Тогда всякая полная подматрица матрицы T определяет в заданном двудольном графе полный двудольный подграф, т. е. биклику, а максимальная полная подматрица — максимальную биклику этого графа. К поиску максимальных биклик сводятся следующие теоретико-графовые задачи:

- в заданном двудольном графе найти все максимальные биклики;
- для заданного двудольного графа найти наименьшее покрытие всех рёбер максимальными бикликами;
- в заданном двудольном графе найти наибольшую биклику;
- для заданного двудольного графа найти наименьшее бикликовое разбиение множества его вершин;
- является ли заданный двудольный граф (k, l)-редким? По определению такой двудольный граф не содержит биклик размера $k \times l$.

Все эти задачи относятся к классу #P-полных или NP-полных задач [10, 11]. Известно, что в общем случае число максимальных биклик графа экспоненциально зависит от числа вершин [12 – 14]. Доказано, что двудольный граф на n вершинах может содержать до $2^{n/2} \approx 1{,}41^n$ максимальных биклик [15, 16]. К родственным можно также отнести перечислительные задачи, связанные с нахождением всех неприводимых покрытий 0,1-матрицы и поиском информативных фрагментов описаний объектов в дискретных процедурах распознавания [17, 18]. Однако большинство известных алгоритмов решения этих задач неприемлемо долго работают на исходных данных большой размерности. Повысить производительность некоторых из них, а также существующих алгоритмов решения рассматриваемой задачи можно путём применения декомпозиционного подхода, излагаемого далее.

3. Метод декомпозиции контекста без потери формальных понятий

Декомпозиционный подход к решению задачи нахождения всех формальных понятий заданного контекста — это сведение её к конечной серии подзадач. Каждая из этих подзадач — уменьшенная копия исходной задачи, которая решается на некоторой части заданного контекста. Процесс декомпозиции направлен на последовательное уменьшение размеров частей контекста. В итоге формируется конечное множество различных частей (в общем случае разного размера и имеющих непустое пересечение). Процесс декомпозиции реализуется итерационно, поскольку рекурсия в подобных случаях более трудоёмка по времени [19]. Для эффективной организации процесса декомпозиции требуется определить правило разложения контекста на части (что является частью и как её выделять в контексте); оценку числа частей, получаемых на каждой итерации разложения; правило остановки процесса разложения. Кроме того, для всякого декомпозиционного метода решения задачи обязательны правила восстановления искомого решения из решений, полученных для подзадач. Полиномиальность по времени процедур разделения исходных данных решаемой задачи на части — требование, при выполнении которого достигается эффект декомпозиции.

Опишем предлагаемый метод декомпозиции формального контекста и докажем, что он позволяет разлагать контекст без потери формальных понятий, а также установим правила эффективной организации процесса декомпозиции.

Пусть $K = (G, M, I)$ — контекст, FC — множество всех его формальных понятий и T — соответствующая ему 0,1-матрица. Контекст $K_1 = (G_1, M_1, I_1)$ назовём частью $K = (G, M, I)$, если $G_1 \subseteq G$, $M_1 \subseteq M$ и для любых $x \in G_1$, $y \in M_1$ отношение $(x, y) \in I_1$ верно тогда и только тогда, когда $(x, y) \in I$. Заметим, что контексту $K_1 = (G_1, M_1, I_1)$ отвечает подматрица матрицы T, у которой удалены строки, соответствующие объектам из $G \setminus G_1$, и столбцы, соответствующие признакам из $M \setminus M_1$. Всякое нетривиальное формальное понятие из FC можно рассматривать в роли части контекста $K = (G, M, I)$. Части $K_1 = (G_1, M_1, I_1)$ и $K_2 = (G_2, M_2, I_2)$ контекста $K = (G, M, I)$ будем считать различными, если $G_1 \neq G_2$ и/или $M_1 \neq M_2$.

Требуется разложить контекст $K = (G, M, I)$ на конечное множество различных частей так, чтобы выполнялись следующие условия:

1) каждая часть содержит, по крайней мере, одно формальное понятие из FC;
2) ни одно формальное понятие из FC не теряется и не возникают новые формальные понятия.

Разложение, удовлетворяющее условиям 1 и 2, назовём «безопасным» относительно формальных понятий. Если 0,1-матрица T полная, то результирующее множество состоит только из одной части, представляющей сам контекст $K = (G, M, I)$, и эта

часть содержит только одно формальное понятие (G, M). Очевидно, что наибольшее число различных частей, на которые можно «безопасно» разложить контекст, равно числу $|FC|$ формальных понятий контекста $K = (G, M, I)$. Поскольку существуют контексты, для которых число формальных понятий экспоненциально зависит от $|G|$ и $|M|$, целесообразно оценить число частей, получаемых на каждой итерации разложения, и определить правило остановки для реализации всего процесса разложения за полиномиальное время.

Пусть $g \in G$ и $m \in M$ — произвольные элементы контекста $K = (G, M, I)$. Пары множеств (g'', g') и (m', m'') образуют формальные понятия, первое из которых назовём объектным, а второе — признаковым формальным понятием контекста $K = (G, M, I)$.

Обозначим $O = \{(g'', g') : g \in G\} \subseteq FC$ множество всех объектных формальных понятий и $S = \{(m', m'') : m \in M\} \subseteq FC$ множество всех признаковых формальных понятий контекста $K = (G, M, I)$.

Утверждение 1. Всякое объектное формальное понятие (g'', g') формального контекста $K = (G, M, I)$ имеет самое большое по размеру содержание среди других формальных понятий, имеющих в объёме объект $g \in G$; признаковое формальное понятие (m', m'') обладает самым большим объёмом среди других формальных понятий, имеющих в содержании признак $m \in M$.

Доказательство. Справедливость утверждения 1 непосредственно следует из свойств оператора $(\cdot)''$ и определения формального понятия. ■

Пара формальных понятий $(g'', g') \in O$, $(m', m'') \in S$ определяет бокс $\omega = (m', g', J)$ как часть контекста $K = (G, M, I)$, если

$$(g'', g') \sqsubseteq (m', m'')\,, \tag{3}$$

что эквивалентно $g'' \subseteq m'$ (или $m'' \subseteq g'$). Про такой бокс будем говорить, что он образован элементами $g \in G$ и $m \in M$. Далее вместо $\omega = (m', g', J)$ будем кратко писать $\omega = (m', g')$ или (m', g').

Утверждение 2. Для всякого формального контекста $K = (G, M, I)$ и любых $(g'', g') \in O$, $(m', m'') \in S$ отношение порядка $(g'', g') \sqsubseteq (m', m'')$ выполняется тогда и только тогда, когда $(g, m) \in I$.

Доказательство. Пусть $(g'', g') \sqsubseteq (m', m'')$. Тогда $g'' \subseteq m'$, $m'' \subseteq g'$. Согласно рефлексивности оператора $(\cdot)''$, имеем $\{g\} \subseteq g'' \subseteq m'$, $\{m\} \subseteq m'' \subseteq g'$. Из (1) следует $(g, m) \in I$. Докажем обратное. Пусть $(g, m) \in I$. Это означает, что $\{g\} \subseteq m'$, $\{m\} \subseteq g'$. В силу монотонности оператора $(\cdot)''$ верны включения $g'' \subseteq (m')''$, $m'' \subseteq (g')''$. Отсюда в силу рефлексивности оператора $(\cdot)''$ и равенств (2) имеем $\{g\} \subseteq g'' \subseteq m'$, $\{m\} \subseteq$ $\subseteq m'' \subseteq g'$. Следовательно, $(g'', g') \sqsubseteq (m', m'')$. ■

Из утверждения 2 следует, что число различных боксов, порождаемых всевозможными элементами формального контекста $K = (G, M, I)$, не превышает веса 0,1-матрицы T, т. е. величины $\|T\|$ — числа единичных элементов этой матрицы. Очевидно, что $1 \leqslant \|T\| \leqslant |G| \cdot |M|$.

Будем говорить, что формальное понятие $(A, B) \in FC$ вложено в бокс (m', g') контекста $K = (G, M, I)$, и записывать $(A, B) \preceq (m', g')$, если $A \subseteq m'$, $B \subseteq g'$. Всякий бокс (m', g') не является пустым, поскольку, согласно (3), он всегда содержит формальные понятия $(g'', g') \in O$ и $(m', m'') \in S$.

Утверждение 3. Всякое нетривиальное формальное понятие (A, B) контекста $K = (G, M, I)$, которое вложено в бокс (m', g'), образованный элементами $g \in G$ и $m \in M$, всегда содержит эти элементы и их замыкания, т. е. если $(A, B) \preceq (m', g')$, то

1) $g \in A$ и $m \in B$;
2) $g'' \subseteq A$ и $m'' \subseteq B$.

Доказательство. Если $(A, B) \preceq (m', g')$, то $A \subseteq m'$, $B \subseteq g'$. В силу антимонотонности отображений $(\cdot)'$ верно $m'' \subseteq A'$, $g'' \subseteq B'$. Для формального понятия (A, B) по определению $A = B'$, $B = A'$. Тогда, $m'' \subseteq B, g'' \subseteq A$. В силу рефлексивности оператора $(\cdot)''$ имеем $\{m\} \subseteq B, \{g\} \subseteq A$. Отсюда следует справедливость обоих высказываний утверждения 3. ■

Согласно утверждению 3, пару (g'', m'') можно рассматривать в качестве типичного представителя не только бокса (m', g'), но и всех формальных понятий контекста $K = (G, M, I)$, вложенных в этот бокс. Это правомерно, поскольку подматрица, соответствующая боксу (m', g'), во всех строках из g'' и всех столбцах из m'' имеет единичные элементы. Соответствие между боксами и формальными понятиями контекста устанавливает следующая теорема.

Теорема 1. Для всякого формального контекста $K = (G, M, I)$, множества FC всех его формальных понятий и любой пары множеств (A, B), $\varnothing \neq A \subseteq G$, $\varnothing \neq B \subseteq M$, справедливы следующие высказывания:

1) если $(A, B) \in FC$, то всегда в $K = (G, M, I)$ существует бокс $\omega = (m', g')$, $g \in G$ и $m \in M$, возможно, не единственный, в который это формальное понятие вложено;
2) если (A, B) — формальное понятие некоторого бокса $\omega = (m', g')$ формального контекста $K = (G, M, I)$, то оно также принадлежит FC.

Доказательство. Пусть (A, B) — произвольное формальное понятие контекста $K = (G, M, I)$ и $\varnothing \neq A \subseteq G$, $\varnothing \neq B \subseteq M$. По определению для него верны равенства

$$(A, B) = (B', A') = (A'', B'')\,. \tag{4}$$

Рассмотрим некоторый объект $g \in A$ и найдём соответствующее ему объектное формальное понятие (g'', g'). Поскольку $\{g\} \subseteq A$, в силу антимонотонности отображений $(\cdot)'$, монотонности оператора $(\cdot)''$ и равенств (4) справедливы отношения

$$A' \subseteq g', \quad g'' \subseteq A'' = A. \tag{5}$$

Аналогично для произвольного признака $m \in B$ и признакового формального понятия (m', m'') верны отношения

$$B' \subseteq m', \quad m'' \subseteq B'' = B. \tag{6}$$

Из (4)–(6) вытекает, что $g'' \subseteq m'$ и $m'' \subseteq g'$. Следовательно, пара формальных понятий (g'', g') и (m', m'') определяет бокс $\omega = (m', g')$. Кроме того, $A = B' \subseteq m'$, $B = A' \subseteq$ $\subseteq g'$. Это означает, что формальное понятие (A, B) вложено в бокс $\omega = (m', g')$. Если выбрать другой объект из A и/или другой признак из B, то получим тот же самый бокс или, возможно, другой бокс, содержащий формальное понятие (A, B). Первое высказывание теоремы 1 доказано.

Докажем второе высказывание. Пусть (A, B) — формальное понятие некоторого бокса $\omega = (m', g')$ как части контекста $K = (G, M, I)$. Далее результаты отображений $(\cdot)'$, вычисленные для $\omega = (m', g')$, а не контекста $K = (G, M, I)$ в целом, будем

отмечать символом ω в нижнем индексе. В этих обозначениях имеем

$$A = B'_\omega \subseteq m', \quad B = A'_\omega \subseteq g'. \tag{7}$$

Отношения (7) отражают вложенность понятия (A, B) в бокс $\omega = (m', g')$. Если понятие (A, B) совпадает с объектным (g'', g') или признаковым формальным понятием (m', m''), по которым образован бокс $\omega = (m', g')$, то второе высказывание тривиальным образом выполняется.

Пусть формальное понятие (A, B) отлично от (g'', g') и (m', m'') и для него верны отношения (7). Требуется показать, что объём и содержание формального понятия (A, B) не могут выйти за границы бокса $\omega = (m', g')$ при вычислении результатов отображений $(\cdot)'$ применительно к контексту $K = (G, M, I)$, т. е. обязательно верны отношения

$$B'_\omega = B' \subseteq m', \quad A'_\omega = A' \subseteq g'. \tag{8}$$

Заметим, что по утверждению 3 всегда $g \in A$ и $m \in B$. Если предположить, что (8) не выполняются, например $m' \subset B'$, то это будет противоречить утверждению 1, согласно которому формальное понятие (m', m'') обладает самым большим объёмом среди других формальных понятий, имеющих в содержании признак $m \in M$. Справедливость (8) означает, что (A, B) является не только формальным понятием бокса $\omega = (m', g')$, но и формальным понятием исходного контекста $K = (G, M, I)$. ■

Согласно теореме 1, разложение контекста $K = (G, M, I)$ на боксы является «безопасным» для любого формального понятия из FC. В теореме 1 исключены случаи, когда FC содержит хотя бы одно из тривиальных формальных понятий $(G, \varnothing)$, $(\varnothing, M)$. Поскольку всегда верны отношения

$$(\varnothing, M) \sqsubseteq (G, \varnothing), \quad (\varnothing, M) \sqsubseteq (G, G'), \quad (M', M) \sqsubseteq (G, \varnothing),$$

контекст $K = (G, M, I)$ можно рассматривать как бокс (G, M). Следовательно, даже в этих исключительных случаях каждый бокс содержит по крайней мере одно формальное понятие из FC, при этом ни одно формальное понятие из FC не теряется.

Из теоремы 1 вытекает важное практическое следствие: искомое множество FC может быть восстановлено путём объединения множеств формальных понятий, выявленных в боксах контекста $K = (G, M, I)$.

Очевидно, что процесс разложения заданного контекста на боксы может быть организован итерационно, поскольку каждый выявленный на первой итерации бокс можно рассматривать в качестве исходного контекста и вновь подвергать декомпозиции. Оценку числа боксов, получаемых на каждой итерации разложения, устанавливает утверждение 2. Определим правила остановки итерационного процесса разложения. Для этого введём понятие плотности бокса.

Пусть $|m'| \cdot |g'|$ — размер бокса (m', g'), а $\|(m', g')\|$ — число его единичных элементов. Плотностью бокса (m', g') назовём величину

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

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

Утверждение 4. Если бокс (m', g'), образованный элементами $g \in G$ и $m \in M$, имеет плотность $\sigma(m', g') = 1$, то $g'' = m'$, $m'' = g'$.

Доказательство. Поскольку $\sigma(m', g') = 1$, то $|m'| \cdot |g'| = \|(m', g')\|$. Это означает, что для любого объекта $g \in m'$ и для всякого признака $m \in g'$ верно $(g, m) \in I$. Отсюда $g'' = m'$, $m'' = g'$. ■

Утверждение 5. Всякий бокс (m', g') с плотностью $\sigma(m', g') = 1$ содержит ровно одно нетривиальное формальное понятие (A, B) контекста $K = (G, M, I)$, совпадающее с ним, т. е. $A = m'$ и $B = g'$.

Доказательство. Пусть $(A, B) \preceq (m', g')$, тогда $A \subseteq m'$, $B \subseteq g'$. Из утверждения 4 следует, что $A \subseteq m' = g''$ и $B \subseteq g' = m''$, а значит, $A \subseteq g''$ и $B \subseteq m''$. Между тем по утверждению 3 верны обратные включения $g'' \subseteq A$ и $m'' \subseteq B$. Следовательно, $A = m'$, $B = g'$. ■

Из утверждения 5 следует, что бокс (m', g') с плотностью 1 вырождается в нетривиальное формальное понятие и не подлежит дальнейшему разложению.

Заметим, что время построения одного бокса составляет $\mathrm{O}(|G| \cdot |M|)$. Согласно утверждению 2, число боксов, возникающих на каждой отдельной итерации процесса декомпозиции, сопоставимо с $\mathrm{O}(|G| \cdot |M|)$. Если ограничить число итераций некоторой константой, то процесс разложения исходного контекста на боксы можно осуществить за полиномиальное время. Дополнительно можно установить ограничение на плотность формируемых боксов.

На практике число боксов, возникающих на каждой отдельной итерации процесса декомпозиции, в ряде случаев может быть уменьшено за счёт удаления вложенных и кратных боксов. Рассмотрим для контекста $K = (G, M, I)$ множество боксов

$$\Omega = \left\{\omega_1, \omega_2, \ldots, \omega_{\|T\|}\right\},$$

где $\omega_i = (m_i', g_i')$, $i = 1, 2, \ldots, \|T\|$. Будем говорить, что бокс $\omega_1 = (m_1', g_1')$ вложен в бокс $\omega_2 = (m_2', g_2')$, и писать $\omega_1 \preceq \omega_2$, если верны теоретико-множественные включения

$$m_1' \subseteq m_2', \quad g_1' \subseteq g_2'.$$

При $m_1' = m_2'$ и $g_1' = g_2'$ боксы ω_1 и ω_2 назовём кратными. Будем считать, что боксы ω_1 и ω_2 сравнимы между собой, если $\omega_1 \preceq \omega_2$ или $\omega_2 \preceq \omega_1$, иначе несравнимы. Таким образом, множество Ω частично упорядочено относительно введённого отношения порядка. С учётом теоремы 1 справедливо

Следствие 1. Для любых $\omega_1, \omega_2 \in \Omega$, таких, что $\omega_1 \preceq \omega_2$, все формальные понятия бокса ω_1 также являются формальными понятиями бокса ω_2 и контекста $K = (G, M, I)$.

Известно, что в частично упорядоченном множестве всегда можно найти взаимно непересекающиеся цепи [20]. Непустое подмножество $\{\omega_{i1}, \omega_{i2}, \ldots, \omega_{il}\}$ множества Ω является цепью, если все элементы этого подмножества попарно сравнимы между собой и линейно упорядочены: $\omega_{i1} \preceq \omega_{i2} \preceq \ldots \preceq \omega_{il}$. Элемент ω_{il} называется максимальным элементом, а величина l — длиной этой цепи. Цепь называется максимальной, если её объединение с любым не принадлежащим ей элементом цепью не является. Две цепи называются взаимно непересекающимися, если они не содержат общих элементов. Число максимальных взаимно непересекающихся цепей и длина самой длинной такой цепи определяются теоремой Дилоурса [20]. Существует алгоритм построения взаимно непересекающихся цепей частично упорядоченного множества, основанный на вычислении максимального паросочетания двудольного графа. В работе [20] доказано, что

время выполнения данного алгоритма полиномиально относительно мощности исходного частично упорядоченного множества и что построенные цепи максимальные и взаимно непересекающиеся.

Согласно следствию 1, максимальный элемент всякой цепи сохраняет все формальные понятия остальных элементов этой цепи. Данные элементы могут быть удалены и тем самым уменьшено число боксов, получаемых на каждой отдельной итерации разложения. Существуют случаи, когда указанный приём не даёт эффекта, например, когда все элементы множества Ω несравнимы между собой или когда множество Ω линейно упорядочено. Однако эти случаи крайне редки для реальных контекстов.

4. Вычислительные эксперименты

Для оценки результативности предложенного метода декомпозиции формального контекста были выполнены вычислительные эксперименты. Эксперименты проводились с помощью программы FCACorpus, базирующейся на алгоритме Close-by-One нахождения всех формальных понятий [21]. Использовались формальные контексты, описывающие коллекции тувинских текстов. Для каждого контекста $K = (G, M, I)$ осуществлялось нахождение множества FC без разложения и с разложением на боксы. Результаты приведены в таблице, где $|G|$ — количество объектов; $|M|$ — количество признаков исходного контекста $K = (G, M, I)$; $\|T\|$ — вес матрицы, соответствующей этому контексту; $|FC|$ — число найденных формальных понятий; N — количество образованных боксов: t — время выполнения программы. Эксперименты выполнялись на компьютере с процессором Intel Core i7-720QM Processor (6M Cache, 1.60 GHz) и ОЗУ размером 4 ГБ.

Результаты экспериментов

Случаи	$\|G\|$	$\|M\|$	$\|\|T\|\|$	$\|FC\|$	N	t, мс
Без разложения на боксы	100	10	480	592	–	788
С разложением на боксы				592	470	420
Без разложения на боксы	500	20	5100	5780	–	79449
С разложением на боксы				5780	5080	13253
Без разложения на боксы	1000	30	14700	14506	–	394520
С разложением на боксы				14506	13876	180144

Как видно из таблицы, значения $|FC|$ во всех случаях (без разложения и с разложением на боксы) полностью совпадают. Это подтверждает «безопасность» разложения контекста на боксы относительно формальных понятий. Число N боксов, образованных при разложении контекста, не превышает величины $\|T\|$, что свидетельствует о правильности утверждения 2. Эксперименты показывают, что применение предложенного метода декомпозиции даёт значительный выигрыш по времени: время выполнения программы FCACorpus при разложении контекста на боксы уменьшается в несколько раз.

Заключение

Представленный метод декомпозиции позволяет повысить производительность алгоритмов решения задачи нахождения всех формальных понятий и применять их для предметных областей, описываемых контекстами большой размерности. Данный метод применим также для теоретико-графовых задач, связанных с нахождением биклик в двудольном графе. Возможны другие методы разложения формального контекста на части, однако они неизменно должны быть «безопасными» относительно формальных понятий.

ЛИТЕРАТУРА

1. *Биркгоф Г.* Теория решеток. М.: Наука, 1984. 568 с.
2. *Ganter B. and Wille R.* Formal Concept Analyses: Mathematical Foundations. Springer Science and Business Media, 2012. 314 p.
3. *Ganter B. and Obiedkov S. A.* Conceptual Exploration. Berlin, Heidelberg: Springer, 2016. 315 p.
4. *Kuznetsov S. O. and Obiedkov S. A.* Comparing Performance of Algorithms for Generating Concept Lattices // J. Experimental and Theoretical Artificial Intelligence. 2002. V. 14. No. 2. P. 189–216.
5. *Simon A. A.* Best-of-Breed approach for designing a fast algorithm for computing fixpoints of Galois Connections // Inform. Sci. 2015. V. 295. No. 2. P. 633–649.
6. *Aslanyan L., Alipour D., and Heidari M.* Comparative analysis of attack graphs // Mathem. Problems of Computer Sci. 2013. No. 40. P. 85–95.
7. *Heydari M., Morales L., Shields C. O., and Sudborough I. H.* Computing Cross Associations for Attack Graphs and other Applications // Proc. 40th Ann. Hawaii Intern. Conf. on System Sciences. Big Island, Hawaii, 2007. P. 270.
8. *Li J., Liu G., Li H., and Wong L.* Maximal biclique subgraphs and closed pattern pairs of the adjacency matrix: A one-to-one correspondence and mining algorithms // J. IEEE Trans. Knowledge and Data Engineering. 2007. No. 19. P. 1625–1637.
9. *Bein D., Morales L., Bein W., et al.* Clustering and the biclique partition problem // Proc. 41st Ann. Hawaii Intern. Conf. on System Sciences. Big Island, Hawaii, 2008. P. 475–483.
10. *Гэри М., Джонсон Д.* Вычислительные машины и труднорешаемые задачи. М.: Мир, 1982. 416 с.
11. *Дугинов О. И.* Сложность задач покрытия графа наименьшим числом полных двудольных графов // Труды Института математики. 2014. Т. 22. Вып. 1. С. 51–69.
12. *Prisner E.* Bicliques in graphs I: bounds on their number // Combinatorica. 2000. V. 20. P. 109–117.
13. *Wood D. R.* On the maximum number of cliques in a graph // Graphs and Combinatorics. 2007. No. 23. P. 1–16.
14. *Moon J. W. and Moser L.* On cliques in graphs // J. Mathematics. 1965. No. 3. P. 23–28.
15. *Pottosina S., Pottosin Y., and Sedliak B.* Finding maximal complete bipartite subgraphs in a graph // J. Appl. Math. 2008. V. 1. No. 1. P. 75–81.
16. *Vania M. F. Dias, Celina M. H. de Figueiredo, and Jayme L. S.* Generating bicliques of a graph in lexicographic order // Theor. Comput. Sci. 2005. No. 337. P. 240–248.
17. *Дюкова Е. В., Журавлев Ю. И.* Дискретный анализ признаковых описаний в задачах распознавания большой размерности // Журн. вычислит. матем. и матем. физики. 2000. Т. 40. № 8. С.1264–1278.
18. *Дюкова Е. В., Инякин А. С.* О процедурах классификации, основанных на построении покрытий классов // Журн. вычислит. матем. и матем. физики. 2003. Т. 43. № 12. С.1884–1895.
19. *Быкова В. В.* Математические методы анализа рекурсивных алгоритмов // Журн. Сибирского федерального университета. Математика и физика. 2008. Т. 3. № 1. С. 372–384.
20. *Harzheim E.* Ordered Sets. New York: Springer, 2005. 390 p.
21. *Монгуш Ч. М., Быкова В. В.* Программа FCACorpus концептуального моделирования тувинских текстов методами анализа формальных понятий. Свидетельство о государственной регистрации программы для ЭВМ № 2018618907, выдано Федеральной службой по интеллектуальной собственности РФ, 2018.

REFERENCES

1. *Birkhoff G.* Lattice Theory. AMS, Providence, 1967. 423 p.
2. *Ganter B. and Wille R.* Formal Concept Analyses: Mathematical Foundations. Springer Science and Business Media, 2012. 314 p.
3. *Ganter B. and Obiedkov S. A.* Conceptual Exploration. Berlin, Heidelberg, Springer, 2016. 315 p.
4. *Kuznetsov S. O. and Obiedkov S. A.* Comparing performance of algorithms for generating concept lattices. J. Experimental and Theoretical Artificial Intelligence, 2002, vol. 14, no. 2, pp. 189–216.
5. *Simon A. A.* Best-of-Breed approach for designing a fast algorithm for computing fixpoints of Galois Connections. Information Sci., 2015, vol. 295, no. 2, pp. 633–649.
6. *Aslanyan L., Alipour D., and Heidari M.* Comparative analysis of attack graphs. Mathem. Problems of Computer Sci., 2013, no. 40, pp. 85–95.
7. *Heydari M., Morales L., Shields C. O., and Sudborough I. H.* Computing cross associations for attack graphs and other applications. Proc. 40th Ann. Hawaii Intern. Conf. on System Sciences, Big Island, Hawaii, 2007, pp. 270.
8. *Li J., Liu G., Li H., and Wong L.* Maximal biclique subgraphs and closed pattern pairs of the adjacency matrix: A one-to-one correspondence and mining algorithms. J. IEEE Trans. Knowledge and Data Engineering, 2007, no. 19, pp. 1625–1637.
9. *Bein D., Morales L., Bein W., et al.* Clustering and the biclique partition problem. Proc. 41st Ann. Hawaii Intern. Conf. on System Sciences, Big Island, Hawaii, 2008, pp. 475–483.
10. *Garey M. and Johnson D.* Computers and Intractability. N.Y., Freeman and Co, 1979. 340 p.
11. *Duginov O. I.* Slozhnost' zadach pokrytiya grafa naimen'shim chislom polnykh dvudol'nykh grafov [The complexity of the problems of covering a graph with the smallest number of complete bipartite graphs]. Proc. Institute of Mathematics, 2014, vol. 22, iss. 1, pp. 51–69. (in Russian)
12. *Prisner E.* Bicliques in graphs I: bounds on their number. Combinatorica, 2000, vol. 20, pp. 109–117.
13. *Wood D. R.* On the maximum number of cliques in a graph. Graphs and Combinatorics, 2007, no. 23, pp. 1–16.
14. *Moon J. W. and Moser L.* On cliques in graphs. J. Mathematics, 1965, no. 3, pp. 23–28.
15. *Pottosina S., Pottosin Y., and Sedliak B.* Finding maximal complete bipartite subgraphs in a graph. J. Appl. Math., 2008, vol. 1, no. 1, pp. 75–81.
16. *Vania M. F. Dias, Celina M. H. de Figueiredo, and Jayme L. S.* Generating bicliques of a graph in lexicographic order. Theoretical Computer Science, 2005, no. 337, pp. 240–248.
17. *Dyukova E. V. and Zhuravlev Yu. I.* Diskretnyy analiz priznakovykh opisaniy v zadachakh ras-poznavaniya bol'shoy razmernosti [Discrete analysis of feature descriptions in high-dimensional recognition tasks]. J. Comput. Mathem. and Mathem. Physics, 2000, vol. 40, no. 8, pp. 1264–1278. (in Russian)
18. *Dyukova E. V. and Inyakin A. S.* O protsedurakh klassifikatsii, osnovannykh na postroyenii pokrytiy klassov [About classification procedures based on classroom construction]. J. Comput. Mathem. and Mathem. Physics, 2003, vol. 43, no. 12, pp.1884–1895. (in Russian)
19. *Bykova V. V.* Matematicheskiye metody analiza rekursivnykh algoritmov [Mathematical methods for analyzing recursive algorithms]. J. Siberian Federal University. Mathematics and Physics, 2008, vol. 3, no. 1, pp. 372–384. (in Russian)
20. *Harzheim E.* Ordered Sets. New York, Springer, 2005. 390 p.
21. *Mongush Ch. M. and Bykova V. V.* Programma FCACorpus kontseptual'nogo modelirovaniya tuvinskikh tekstov metodami analiza formal'nykh ponyatiy [The program FCACorpus of

conceptual modeling of Tuvan texts by the methods of the formal concepts analysis]. Certificate of state registration of computer programs no. 2018618907, issued by the Federal Service for Intellectual Property of the Russian Federation, 2018. (in Russian)