Инженерный вестник Дона, №11 (2023)
ivdon.ru/ru/magazine/archive/n11y2023/8840

Алгоритм фрагментации и дефрагментации формальных контекстов

Ч.М. Монгуш

Тувинский государственный университет, Кызыл, Республика Тыва

Аннотация: Рассматривается комбинаторная задача – задача нахождения множества всех формальных понятий формального контекста. Вычислительная сложность задачи состоит в том, что число формальных понятий экспоненциально зависит от размера исходного формального контекста. В статье для решения данной задачи приводится алгоритм фрагментации и дефрагментации формального контекста, основанный на методе декомпозиции формального контекста на фрагменты. Суть метода заключается в том, что исходный формальный контекст разделяется на различные фрагменты. Фрагменты имеют разные размеры и непустое пересечение. Каждый фрагмент в дальнейшем рассматривается, как формальный контекст и вновь может подвергаться к декомпозицию. В итоге формируется конечное множество фрагментов. Затем в каждом фрагменте находятся формальные понятия и объединяются, образуя искомое множество всех формальных понятий формального контекста. Метод является «неискажающим»: при разделении контекста на фрагменты не образуются новые формальные понятия и не теряются искомые понятия. Приводятся результаты вычислительных экспериментов, показывающие эффективность разработанного алгоритма.
Ключевые слова: анализ формальных понятий, алгоритм фрагментации, формальный контекст, объектно-признаковая таблица, комбинаторная задача, задача нахождения множества всех формальных понятий.

Введение

Формальный контекст в алгебраическом методе Р. Вилле и Б. Гантера, который называется анализ формальных понятий, является объектно-признаковой таблицей [1, 2]. Каждая строка таблицы – признаковое описание некоторого объекта. В свою очередь формальный контекст можно представить в виде (0,1) -матрицы. Тогда формальное понятие формального контекста – максимально полная подматрица объектно-признаковой матрицы [3–5]. Простым способом нахождения всех формальных понятий контекста является перебор всех возможных подмножеств множества объектов или признаков. Поэтому задача нахождения множества всех формальных понятий относится к комбинаторным задачам [2]. В статье для решения данной задачи приводится алгоритм фрагментации и дефрагментации формального контекста, основанный на методе декомпозиции формального контекста на фрагменты.

Основные понятия

Рассмотрим пример формального контекста $\digamma = \langle O,\ P,\ J\rangle$, где множество объектов $O = \{o1, o2, o3, o4, o5\}$, набор свойств $P = \{p_1, p_2, p_3, p_4, p_5\}$, матрица инцидентности $J \subseteq O{\times}P$, представленная в табл. 1. Запись $(o, p){\in}J$ означает, что объекту o присущ свойство p. Эта запись в таблице обозначается «+».

Таблица № 1

Формальный контекст $\digamma = \langle O, P, J\rangle$

	p_1	p_2	p_3	p_4	p_5
o1	+		+	+	
o2		+	+		+
o3	+	+	+		+
o4		+			+
o5	+	+	+		+

Определим для элементов формального контекста $F = \langle O,\ P,\ J\rangle$ отображения Галуа $(\cdot)'$: $o' = \{p{\in}P :\ (o, p){\in}J\ \}$, $p' = \{o \in O :\ (o, p){\in}J\}$.
По определению o' – множество признаков, присущих объекту o, а p' – объекты, обладающие свойством p. Пусть $\{o1\} \in O$, $\{p_1\}{\in}P$ и определим для них отображения Галуа $(\cdot)'$: $(\{o1\})' = \{p_1, p_3, p_4\}$, $(\{p_1\})' = \{o1, o3, o5\}$.

Обобщим применение отображения на множества $X \subseteq O$ и $Y \subseteq P$:

$X' = \cap_{\,o \in X}\, o' = \{p \in P \mid \forall o \in X\ (o, p) \in J\}$,

$Y' = \cap_{\,p \in Y}\, p' = \{o \in O \mid \forall p \in Y\ (o, p) \in J\}$.

Аналогично, для $X = \{o1, o3, o5\} \subseteq O$ и $Y = \{p_1, p_3, p_4\} \subseteq P$:

$X' = (\{o1, o3, o5\})' = \{p_1, p_3\}$,

$Y' = (\{p_1, p_3, p_4\})' = \{\ o1\}$.

В этом случае X' – множество признаков, характерных всем объектам из X, а Y' – совокупность объектов, обладающих всеми признаками из Y.

Целесообразно положить, что пустому множеству объектов характерны все свойства P, а всякий объект из O обладает пустым множеством признаков: $(\varnothing)' = P$ и $(\varnothing)' = O$.

Двойное применение отображения Галуа устанавливает оператор замыкания $()''$. Он вычисляется по формуле:

$$Y'' = (Y')' = \begin{cases} \left(\bigcap_{p\in Y} p'\right)' = \bigcap_{o\in Y'} o', \textit{если } Y' \neq \varnothing \\ P, \textit{если } \quad Y' = \varnothing \end{cases} \qquad (1)$$

Если $Y \subseteq P = Y''$, то множество Y называется замкнутым множеством. Аналогично и для $X \subseteq O$. Например, пусть X=$\{o1, o3, o5\} \subseteq O$, тогда

$X'' = (X')' = (\{o1, o3, o5\}')' = (\{p_1, p_3\})' = \{ o1, o3, o5\} = X$.

Отсюда следует, что X – замкнутое множество.

Формальным понятием называется пара множеств (X, Y), где $X \subseteq O$ и $Y \subseteq P$, если выполняется $X' = Y$ и $Y' = X$ [6]. Другими словами, множества X, Y образуют формальное понятие (X, Y) в формальном контексте $\mathcal{F} = \langle O, P, J\rangle$ тогда и только тогда, когда:

$$X = X'' \text{ и } Y = Y''. \qquad (2)$$

Найдем все формальные понятия контекста $\mathcal{F} = \langle O, P, J\rangle$ простым способом. Для этого необходимо: подобрать всевозможные подмножества множества O или P; для каждого подмножества множества O или P вычислить замыкание по формуле (1); проверить полученный результат на соответствие условию (2).

В итоге получим следующее множество всех формальных понятий формального контекста:

$(O, \varnothing)$, $(o1, p_1\ p_3\ p_4)$, $(o1\ o3\ o5, p_1\ p_3)$, $(o1\ o2\ o3\ o5, p_3)$, $(o3\ o5, p_1\ p_2\ p_3\ p_5)$, $(o2\ o3\ o5, p_2\ p_3\ p_5)$, $(o2\ o3\ o4\ o5, p_2\ p_5)$, $(\varnothing, P)$. (3)

Если представить формальное понятие в (0,1) -матрице, то это будет максимально полная подматрица. В таблице 2 показано формальное понятие ($o2$ $o3$ $o4$ $o5$, p_2 p_5), закрашенное серым цветом.

Таблица № 2

Формальное понятие ($o2$ $o3$ $o4$ $o5$, p_2 p_5)

	p_1	p_2	p_3	p_4	p_5
$o1$	+		+	+	
$o2$		+	+		+
$o3$	+	+	+		+
$o4$		+			+
$o5$	+	+	+		+

В настоящее время для вычисления множества формальных понятий контекста существуют много алгоритмов и программных комплексов. Наиболее известные из них – Concept Explorer, ToscanaJ, Oblicia, Lattice Minner, OpenFCA, FCART. Их подробный обзор представлен в работах [7, 8].

В данной работе представлен алгоритм нахождения множества всех формальных понятий, основанный на методе фрагментации и дефрагментации формального контекста. Подробное описание метода показано в работах [3, 6].

Алгоритм фрагментации и дефрагментации формального контекста

Суть метода заключается в том, что исходный формальный контекст разбивается на различные фрагменты. Фрагменты имеют разные размеры и непустое пересечение. Каждый фрагмент в дальнейшем рассматривается, как формальный контекст и снова подвергается разложению. В итоге формируется конечное множество фрагментов (рис. 1).

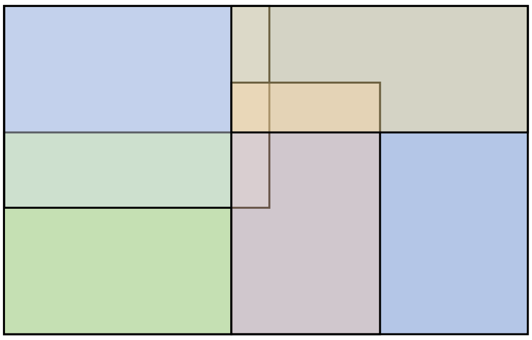

Рис. 1. – Фрагментация формального контекста

Затем в каждом фрагменте находятся формальные понятия путем известного алгоритма Close-by-One. Полученные формальные понятия объединяются и образуется искомое множество всех формальных понятий формального контекста. В работе [3] доказано, что метод фрагментации и дефрагментации формального контекста является «неискажающим», т.е. каждый фрагмент содержит хотя бы одно формальное понятие, при разложении контекста не исчезают искомые формальные понятия и не появляются новые понятия. Процесс разложения контекста производится итерационно. Для остановки данного процесса необходимо указать количество итераций, а также дополнительно можно ввести порог на плотность фрагментов. На основе данного метода был разработан алгоритм, который приведен ниже.

На вход алгоритм получает формальный контекст $\mathcal{F} = \langle O,\ P,\ J\rangle$ и количество итераций. В Ω_1 хранятся фрагменты, которые подлежат дальнейшему декомпозицию, Ω_2 – фрагменты, порожденные в формальное понятие, H_1, H_2 – типичные представители фрагментов, входящих в Ω_1 и Ω_2 соответственно.

Выполнение алгоритма рассмотрим на примере формального контекста, представленного в таблице 1. Пусть $k = 1$, $o \in O$, $p \in P$.

Алгоритм. Нахождение всех формальных понятий формального контекста

Вход: исходный формальный контекст $\mathcal{F} = \langle O, P, J\rangle$, k – количество итераций

begin
$\Omega_1 = \mathcal{F} = \langle O, P, J\rangle$
$\Omega_2 = \varnothing$
$H_1 = (O'', P'')$
$H_2 = \varnothing$
while $(k \neq 0 \;\&\; \Omega_1 \neq \varnothing)$ **do**
  for all $\omega \in \Omega_1$ **do**
    if $\sigma(\omega) \neq 1$ **then**
      Boxes(ω)
    else
      $\Omega_2 = \Omega_2 \cup \omega$
      $H_2 = H_2 \cup H_1$
    end if
  end for
  Delete($\Omega_1 \cup \Omega_2$, $H_1 \cup H_2$)
  if $\Omega_1 \neq \varnothing$ **then**
    SearchChains(Ω_1, H_1)
  end if
  $k = k - 1$
end while
$\Omega = \Omega_1 \cup \Omega_2$
$H = H_1 \cup H_2$
end

Выход: Ω - множество фрагментов, H – множество типичных представителей фрагментов

Для нахождения фрагментов формального контекста $\mathcal{F} = \langle O, P, J\rangle$ необходимо:

1. Вычислить плотность фрагмента (формального контекста) по формуле:

$$\sigma = \|(O, P, J)\| / (|O| \cdot |P|),$$

где $\|O, P, J\|$ – количество «+» элементов в матрице, $(|O| \cdot |P|)$ – размер контекста. Если плотность фрагмента равна 1, то данный фрагмент является формальным понятием и не подлежит разложению, иначе фрагмент подлежит декомпозиции.

2. Найти объектные (o'', o') и признаковые (p', p'') формальные понятия контекста.

3. Выявить фрагмент $\Omega = (p', o')$ по формуле:

$$(o'', o') \le (p', p''), \text{ что эививалентно } o'' \subseteq p' \text{ (или } p'' \subseteq o'). \qquad (4)$$

Будем говорить, что фрагмент Ω образован элементами $o \in O$, $p \in P$. Тогда число фрагментов контекста $F = \langle O, P, J \rangle$, формируемых всеми элементами контекста, не превышает веса (0,1)-матрицы. Типичным представителем фрагмента называется пара (o'', p''). Она также является типичным представителем формальных понятий, вложенных в данный фрагмент. Так как подматрица, соответствующая фрагменту $\Omega = (p', o')$, в строках из o'' и столбцах p'' имеет «+».

Плотность формального контекста, представленного в таблице 1, равна $\sigma(F) = 0{,}64 \neq 1$. Следовательно, формальный контекст можно разложить на фрагменты. Количество фрагментов контекста не должен превышать 16.

Зафиксируем объект $o = \{o1\}$, признак $p = \{p_1\}$. Тогда объектное понятие имеет вид $(o'', o') = (o1, p_1p_3p_4)$, признаковое понятие $(p', p'') = (o1o3o5, p_1p_3)$. Нетрудно увидеть, что для них условие (4) выполняется:

$$(o1, p_1p_3p_4) \le (o1o3o5, p_1p_3), \text{ т.е. } \{o1\} \subseteq \{o1o3o5\} \text{ и } \{p_1p_3\} \subseteq \{p_1p_3p_4\}.$$

Следовательно, образуется фрагмент $\Omega = (o1o3o5, p_1p_3p_4)$, типичным представителем которого является $(o1, p_1p_3)$. В таблице 3 показан данный фрагмент и выделен типичный представитель.

Таблица № 3

Фрагмент $\Omega = (o1o3o5, p_1p_3p_4)$

	p1	$\boldsymbol{p_3}$	$\boldsymbol{p_4}$	$\boldsymbol{p_5}$	$\boldsymbol{p_2}$
o1	+	+	+		
***o*3**	+	+		+	+
***o*5**	+	+		+	+
***o*2**		+		+	+
***o*4**				+	+

Аналогичным образом находим остальные фрагменты. Кратные и совпадающие с исходным контекстом фрагменты удаляются с помощью функции Delete. После этого, из 16 фрагментов остаются 9 фрагментов.

Выявление вложенных фрагментов и поиск максимального из них осуществляется в процедуре SearchChains путем построения взаимно непересекающихся цепей частично упорядоченного множества фрагментов согласно теореме Дилоурса [6]. В результате работы процедуры SearchChains остаются 3 фрагмента, которые представлены в таблице 4.

Таблица № 4

Фрагменты и их типичные представители

Фрагмента контекста	**Типичные представители**	**Плотность**	**Формальные понятия, вложенные в фрагмент**
$(o1o2o3o5,\ p_1p_3p_4)$	$(o1,\ p_3)$	0,67	$(o1o3o5,\ p_1p_3)$ $(o1,\ p_1p_3p_4)$ $(o1o2o3o5,\ p_3)$
$(o2\ o3\ o4\ o5,\ p_1p_2p_3p_5)$	$(o3o5,\ p_2p_5)$	0,81	$(o2\ o3\ o4\ o5,\ p_2\ p_5)$ $(o2\ o3\ o5,\ p_2\ p_3p_5)$ $(o3o5,\ p_2\ p_5)$ $(o3\ o5,\ p_1\ p_2\ p_3\ p_5)$
$(o1\ o2\ o3\ o5,\ p_1p_2p_3p_5)$	$(o3o5,\ p_3)$	0,81	$(o1\ o3\ o5,\ p_1\ p_3)$ $(o1\ o2\ o3\ o5,\ p_3)$ $(o2\ o3\ o5,\ p_2\ p_3\ p_5)$ $(o3\ o5,\ p_1\ p_2\ p_3\ p_5)$

После этого в каждом фрагменте находятся формальные понятия с помощью известного алгоритма Close-by-One. В последнем столбце таблицы 4 указаны формальные понятия фрагмента. Затем восстановить искомое множество всех формальных понятий с использованием теоретико-множественной операции объединения. В результате получим:

$(O, \varnothing)$, $(o1,\ p_1\ p_3\ p_4)$, $(o1\ o3\ o5,\ p_1\ p_3)$, $(o1\ o2\ o3\ o5,\ p_3)$, $(o3\ o5,\ p_1\ p_2\ p_3\ p_5)$, $(o2\ o3\ o5,\ p_2\ p_3\ p_5)$, $(o2\ o3\ o4\ o5,\ p_2\ p_5)$, $(\varnothing, P)$.

Полностью совпадает с искомым множеством, приведенным в (3).

Аналогично можно выполнить вторую итерацию, т. е. полученные фрагменты вновь разложить на фрагменты и т.д. Увеличение числа итераций приводит к увеличению числа фрагментов, подлежащих дальнейшему разложению, к уменьшению их размеров, к увеличению плотности формируемых фрагментов.

Время выполнения алгоритма для однократного разложения контекста на фрагменты составляет $O(\sigma(F) \cdot |O|^2 \cdot |P|^2)$. Если количество итераций равно k, то процесс разложения исходного контекста на фрагменты алгоритмом выполняется за время $O(|O|^{2k} \cdot |P|^{2k})$. Поэтому k необходимо фиксировать. В этом случае время выполнения алгоритма будет полиномиальным относительно размера исходного формального контекста.

Представленный алгоритм можно использовать при решении задач обработки естественного языка [9, 10].

Вычислительные эксперименты

Были проведены серии вычислительных экспериментов. Первый эксперимент направлен на результативность приведенного алгоритма. Были рассмотрены формальные контексты, заполненные случайным образом. Оценивалось время, затраченное на нахождение множества всех формальных понятий формального контекста с помощью алгоритма фрагментации и дефрагментации формального контекста и без декомпозиции контекста. Результаты приведены в таблице 5.

Таблица № 5

Результаты первого эксперимента

	Размер формального контекста	Время, мс
С алгоритмом фрагментации	100×20	2450
Формальный контекст не разбивается на фрагменты		150533
С алгоритмом фрагментации	200×30	91200
Формальный контекст не разбивается на фрагменты		810530

Из полученного результата следует, что на нахождение множества всех формальных понятий путем разделения контекста на фрагменты отводится значительно меньше время по сравнению с алгоритмом без разбиения. Чтобы время выполнения алгоритма была полиномиальным, необходимо правильно выбрать число итераций. Для проведения этого эксперимента берется формальный контекст вида $F = \langle O; O;\neq\rangle$. Результаты представлены на рис. 2.

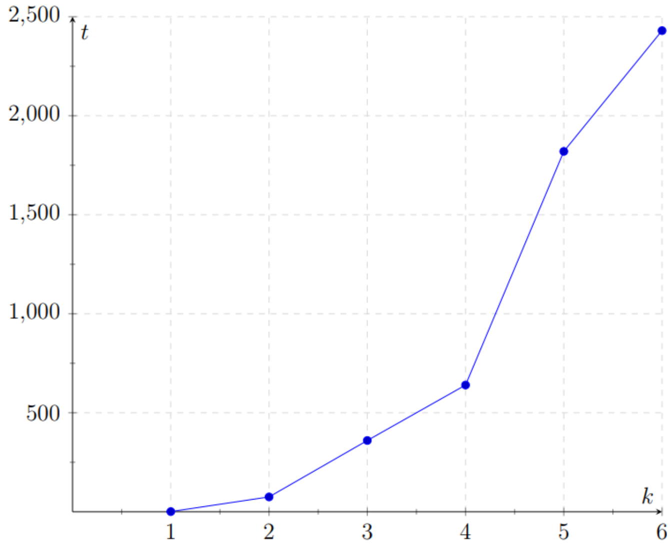

Рис. 2. – Результаты второго эксперимента

Из рисунка видно, что чем больше число итераций, тем больше требуется времени на выполнение алгоритма. Поэтому рекомендуется значение k выбрать маленьким, например, 2, 3 или 4.

Заключение

Таким образом, чтобы снизить вычислительную сложность задачи нахождения всех формальных понятий, в работе предложен алгоритм фрагментации и дефрагментации формального контекста. Показана эффективность предложенного алгоритма при соблюдении рекомендаций.

Исследование выполнено при финансовой поддержке гранта Президента Российской Федерации для молодых ученых – кандидатов наук

МК-2437.2022.1.1 «Разработка математического метода и средств для исследования текстов на тувинском языке».

Литература

1. Ganter B., Wille R. Formal Concept Analyses: Mathematical Foundations. Springer Science and Business Media, 2012. 314 p.

2. Кузнецов С.О. Автоматическое обучение на основе анализа формальных понятий // Автоматика и телемеханика. 2001. № 10. С. 3–27.

3. Быкова В.В., Монгуш Ч.М. Декомпозиционный подход к исследованию формальных контекстов // Прикладная дискретная математика. 2019. № 44. С. 113–126.

4. Белим С.В., Богаченко Н.Ф. Использование решетки формальных понятий для построения ролевой политики разграничения доступа // Информатика и системы управления. 2019. № 1(55). С. 16–28.

5. Kuznetsov S.O., Ganter B., Eklund P.W., Sertkaya B. Machine Learning and Formal Concept Analysis // 2th International Conference on Formal Concept Analysis: proceedings. Berlin Heidelberg: Springer, 2004. Vol. 2961. pp. 287–312.

6. Mongush Ch.M. Bykova V.V. On decomposition of a binary context without losing formal concepts // Journal of Siberian Federal University. Mathematics and Physics. 2019. Vol. 12. № 3. pp. 323–330.

7. Qian T., Wei L., Qi J. Decomposition methods of formal contexts to construct concept lattices // International Journal of Machine Learning and Cybernetics. 2017. Vol. 8. № 1. pp. 95–108.

8. Simon A. A. Best-of-Breed approach for designing a fast algorithm for computing fixpoints of Galois Connections // Information Sciences. 2015. Vol. 295. № 2. pp. 633–649.

9. Монгуш Ч.М. Концептуальное моделирование текстов тувинского языка с использованием методов анализа формальных понятий //

Инженерный вестник Дона, 2022, №11. URL: ivdon.ru/ru/magazine/archive/n11y2022/7995.

10. Королев И.Д. Способ автоматизированного формирования обучающего набора данных для алгоритмов машинного обучения классификации электронных документов // Инженерный вестник Дона, 2023, №10. URL: ivdon.ru/ru/magazine/archive/n10y2023/8726

References

1. Ganter B., Wille R. Formal Concept Analyses: Mathematical Foundations. Springer Science and Business Media, 2012. 314 p.

2. Kuznetsov S.O. Avtomatika i telemekhanika. 2001. №10. pp. 3–27.

3. Bykova V.V., Mongush Ch.M. Prikladnaia diskretnaia matematika. 2019. №44. pp. 113–126.

4. Belim S.V., Bogachenko N.F. Informatika i sistemy upravlenija. 2019. №1 (55). pp. 16–28.

5. Kuznetsov S.O., Ganter B., Eklund P.W., Sertkaya B. 2th International Conference on Formal Concept Analysis: proceedings. Berlin Heidelberg, Springer, 2004, vol. 2961. pp. 287–312.

6. Mongush Ch. M., Bykova V.V. Journal of Siberian Federal University. Mathematics and Physics. 2019. vol. 12. №3. pp. 323–330.

7. Qian T., Wei L., Qi J. International Journal of Machine Learning and Cybernetics. 2017. vol. 8. № 1. pp. 95–108.

8. Simon A. A. Information Sciences. 2015. vol. 295. № 2. pp. 633–649.

9. Mongush Ch.M. Inzhenernyj vestnik Dona, 2022, №11. URL: ivdon.ru/ru/magazine/archive/n11y2022/7995.

10. Korolev I.D. Inzhenernyj vestnik Dona, 2023, №10. URL: ivdon.ru/ru/magazine/archive/n10y2023/8726.