УДК 519.688:004.93
DOI: 10.15827/0236-235X.119.487-495
Дата подачи статьи: 15.06.17
2017. Т. 30. № 3. С. 487–495

АЛГОРИТМЫ КОНЦЕПТУАЛЬНОГО МОДЕЛИРОВАНИЯ И КЛАССИФИКАЦИИ ТЕКСТОВ В КОРПУСЕ ТУВИНСКОГО ЯЗЫКА

В.В. Быкова, д.ф.-м.н., профессор, bykvalen@mail.ru
(Сибирский федеральный университет,
Свободный просп., 79, г. Красноярск, 660041, Россия);
Ч.М. Монгуш, аспирант, преподаватель, mongushchod91@yandex.ru
(Сибирский федеральный университет, Свободный просп., 79, г. Красноярск, 660041, Россия;
Тувинский государственный университет, ул. Ленина, 36, г. Кызыл, 667000, Россия)

Корпус языка – это информационно-лингвистическая система, основанная на собрании оцифрованных текстов некоторого языка. На сегодняшний день корпус тувинского языка включает официально-деловые документы и произведения тувинской художественной литературы.

Работы по расширению корпуса тувинского языка и углублению уровня обработки текстов продолжаются. Они приводят к необходимости решения задач анализа естественно-языковых текстов. Основными из этих задач являются классификация по прецедентам и концептуальное моделирование.

Для их решения в статье используется алгебраический подход, называемый анализом формальных понятий. Предлагаются алгоритмы и программы для построения концептуальной модели коллекции литературных произведений и решения задачи бинарной классификации по прецедентам. Указаны приемы снижения вычислительной сложности рассматриваемых алгоритмов.

В работе представлены результаты вычислительных экспериментов, подтверждающие результативность предложенных приемов по снижению сложности вычислений. Приведены результаты концептуального моделирования и бинарной классификации произведений тувинского фольклора.

Ключевые слова: *корпус языка, анализ формальных понятий, концептуальные модели текстов, алгоритм классификации, алгоритмы снижения размерности контекстов.*

В настоящее время активно создаются корпусы языков народов Российской Федерации для сохранения национального литературного наследия и проведения научных исследований по изучению этих языков. Работа над созданием Национального корпуса тувинского языка ведется сотрудниками, аспирантами и студентами Тувинского государственного и Сибирского федерального университетов [1]. Под корпусом понимается информационно-лингвистическая система, основанная на собрании оцифрованных текстов. Корпус включает в себя различные типы текстов, представленных в языке, а также разметку – информацию о свойствах текстов. В рамках корпусов решаются многие задачи анализа естественно-языковых текстов, возникающие в филологических и лингвистических исследованиях [2]. Основными из них являются классификация по прецедентам и концептуальное моделирование. Классификация по прецедентам, как правило, направлена на установление жанра и автора текста, определение пространственно-временного периода написания произведения. Цель концептуального моделирования – структурное представление знаний, извлеченных из текстов произведений (например, особенности использования языковых клише и диалектных вариантов эпических выражений).

Традиционными подходами для решения задачи классификации являются методы машинного обучения и математической статистики, основанные на количественных мерах близости рассматриваемых текстов [3]. В этом случае, как правило, семантическая составляющая анализируемых текстов не рассматривается. Для построения концептуальных моделей текстов обычно используются концептуальные графы и решетки понятий [3]. Однако существует формализованный подход, известный в литературе как анализ формальных понятий, который позволяет выполнять классификацию, концептуальное моделирование и решать многие другие задачи анализа текстов на основе алгебраической теории решеток [4, 5]. В рамках этого подхода тексты анализируются на семантическом уровне, моделируются и представляются бинарными контекстами, отражающими наличие или отсутствие признаков, характерных для исследуемых произведений. Использование методов анализа формальных понятий позволяет не включать в ПО корпусов пакеты программ, ориентированные на решение отдельных задач анализа данных.

Анализ формальных понятий (АФП) является прикладной ветвью алгебраической теории решеток, в рамках которой возможна формализация терминов «понятие» и «иерархия понятий».

Основные идеи АФП были сформулированы в работе Р. Вилле и Б. Гантера [6] и развиты в исследованиях С.О. Кузнецова, Д.И. Игнатова, С.И. Гурова [5, 7]. В АФП формальные понятия определяются с помощью соответствий Галуа и представляют собой пары множеств вида (объем, содержание). Основным достоинством такого определения является абсолютное соответствие традиционной трактовке термина «понятие», используемого в гуманитарных науках [8].

В статье предлагаются алгоритмы и программы для построения концептуальной модели и решения задачи бинарной классификации по прецедентам с помощью математического аппарата АФП. Указаны приемы снижения вычислительной сложности рассматриваемых алгоритмов. Представлены результаты вычислительных экспериментов, подтверждающие результативность предложенных приемов по снижению сложности вычислений. Приведены результаты концептуального моделирования и бинарной классификации по прецедентам произведений тувинского фольклора.

Основные термины и обозначения

Приведем термины и обозначения, применяемые в АФП [6].

Пусть для некоторой предметной области определены два непустых конечных множества G и M объектов и признаков соответственно (от немецких слов *Gegenstände* – объект, *Merkmale* – признак). Пусть также задано отношение инцидентности между множествами G и M: $I \subseteq G \times M$. Существование в I пары (g, m), $g \in G$ и $m \in M$, означает, что объект g имеет признак m и, наоборот, признак m присущ объекту g. Тройка $K = (G, M, I)$ называется формальным контекстом (или кратко контекстом) для рассматриваемой предметной области. Если множества G и M линейно упорядочены (например лексикографически), то контекст можно однозначно (с точностью до «материальной» природы объектов и признаков) задать бинарной матрицей T, отражающей отношение инцидентности I.

Выберем два произвольных элемента $g \in G$ и $m \in M$. Определим для них отображения φ и ψ:

$\varphi(g) = \{m \in M \mid (g, m) \in I\}$ – множество признаков, присущих объекту $g \in G$;

$\psi(m) = \{g \in G \mid (g, m) \in I\}$ – множество объектов, обладающих признаком $m \in M$.

Отображения φ и ψ можно обобщить на произвольные $A \subseteq G$ и $B \subseteq M$ следующим образом:

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

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

Здесь $\varphi(A)$ – множество признаков, общих для всех объектов из A, а $\psi(B)$ – множество объектов, обладающих всеми признаками из B. При этом считается, что $\varphi(\varnothing) = M$ и $\psi(\varnothing) = G$, то есть пустому множеству объектов присущи все признаки из M, и каждый объект рассматриваемого контекста обладает пустым множеством признаков. Отображения φ и ψ определены так, что для любых $A_1, A_2 \subseteq G$ и $B_1, B_2 \subseteq M$ верны равенства:

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

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

Обычно в анализе формальных понятий для отображений φ и ψ применяется единое обозначение $(\cdot)'$, а приведенные выше формулы для $\varphi(A)$, $\psi(B)$ записываются так:

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

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

Если $g \in G$ и $m \in M$, то обозначения g' и m' служат сокращенной формой записи множеств $\varphi(g) = \{g\}'$ и $\psi(m) = \{m\}'$ соответственно. Отображения «$'$» удовлетворяют свойствам, вытекающим из их определения и вполне реалистичного и постулируемого в анализе данных положения: расширение (сокращение) множества признаков уменьшает (увеличивает) число объектов, обладающих этими признаками. Формально эти свойства можно выразить в виде следующих утверждений.

Утверждение 1. Для всякого контекста $K = (G, M, I)$ и любых $B_1, B_2 \subseteq M$ верны свойства:

– антимонотонность:

если $B_1 \subseteq B_2$, то $(B_2)' \subseteq (B_1)'$;

– экстенсивность:

$B_1 \subseteq (B_1)''$, где $(B_1)'' = ((B_1)')' \subseteq M$.

Утверждение 2. Для всякого контекста $K = (G, M, I)$ и любых $A_1, A_2 \subseteq G$ верны свойства:

– антимонотонность:

если $A_1 \subseteq A_2$, то $(A_2)' \subseteq (A_1)'$;

– экстенсивность:

$A_1 \subseteq (A_1)''$, где $(A_1)'' = ((A_1)')' \subseteq G$.

Множество $(B_1)'' = \varphi(\psi(B_1))$ можно трактовать как набор признаков, которые всегда появляются в объектах контекста $K = (G, M, I)$ вместе с признаками из B_1, причем это множество является наибольшим по включению в пределах этого контекста. Множество $(A_1)'' = \psi(\varphi(A_1))$ можно интерпретировать как наибольшее по включению множество объектов, которые обладают всеми признаками, характерными для объектов A_1. Согласно утверждениям 1 и 2, отображения φ и ψ составляют пару соответствий Галуа между множествами 2^G и 2^M, частично упорядоченными по включению [9]. Здесь традиционно 2^G и 2^M – совокупность всех подмножеств рассматриваемых множеств G и M соответственно. Двойное применение отображения «$'$» определяет оператор замыкания «$''$» на 2^G или 2^M в алгебраическом смысле [9].

Множество признаков $B \subseteq M$, для которого $B = B''$, называется замкнутым в контексте $K = (G, M, I)$. Принято говорить, что множество B'' является замыканием для $B \subseteq M$ в контексте $K = (G, M, I)$. Исходя из (1) и (2), при $B' \neq \varnothing$ замыкание для $B \subseteq M$ можно вычислить по формуле

$$B'' = \cap_{g \in G} \{g' \mid B \subseteq g'\}. \quad (3)$$

Если $B' = \varnothing$, то всегда $B'' = \varphi(\psi(B)) = \varphi(\varnothing) = M$. Важно отметить, что применение формулы (3) позволяет за один просмотр контекста $K = (G, M, I)$ найти замыкание для заданного множества признаков.

Концептуальное моделирование и АФП

Концептуальное моделирование направлено на структурное представление знаний об исследуемой

предметной области. Концептуальная модель как семантическая структурная модель предметной области представляет группы однородных объектов, связанных между собой отношениями. Однородность объектов в данном случае понимается как наличие у них одних и тех же свойств. Эти группы однородных объектов традиционно называют сущностями, концептами или понятиями. Объекты, образующие понятия, могут иметь различную природу. В концептуальных моделях используются не только бинарные отношения, но и отношения произвольной арности. Такая универсальность позволяет применять концептуальное моделирование в самых разных областях, в том числе в анализе текстов. В общем случае описание концептуальной модели содержит перечень взаимосвязанных понятий предметной области вместе с их свойствами и характеристиками.

Для описания концептуальных моделей используются различные средства, но в основном графические (диаграммы «сущность–связь», концептуальные графы, решетки понятий). В системах обработки текстовых данных применяются преимущественно концептуальные графы и решетки [3]. Анализ формальных понятий позволяет не только представить (описать и визуализировать) концептуальную модель, но и построить ее, исходя из формальных контекстов. Формальный контекст является абстрактной моделью исследуемой предметной области, отражающей отношение инцидентности между объектами и их свойствами. Имея в наличии формальный контекст, с помощью математических методов АФП можно выявить основные понятия предметной области, установить между ними отношение частичного порядка и объединить формальные понятия в решетку. Такая решетка понятий – математическое описание концептуальной модели, допускающее ее исследование математическими методами.

Дадим определение решетки формальных понятий. Пусть для некоторой предметной области задан формальный контекст $K = (G, M, I)$, где G – множество объектов; M – множество признаков; I – отношение инцидентности между множествами G и M. Пара (A, B), $A \subseteq G$, $B \subseteq M$, такая, что $A' = B$ и $B' = A$, называется формальным понятием с объемом A и содержанием B. Другими словами, пара множеств (A, B) является формальным понятием для контекста $K = (G, M, I)$ тогда и только тогда, когда $A = A''$ и $B = B''$, то есть когда A, B – замкнутые множества относительно оператора «$''$». Если контекст $K = (G, M, I)$ представлен матрицей T, то формальному понятию (A, B) соответствует ее максимальная подматрица, заполненная единицами. Строки этой подматрицы отвечают элементам из A, а столбцы – элементам из B.

Пусть FC_K – множество всех формальных понятий контекста $K = (G, M, I)$. Введем на FC_K отношение частичного порядка $\sqsubseteq$ следующим образом: $(A_1, B_1) \sqsubseteq (A_2, B_2)$, если $A_1 \subseteq A_2$ (или $B_2 \subseteq B_1$), (4) где $A_1, A_2 \subseteq G$ и $B_1, B_2 \subseteq M$. Заметим, что в высказывании (4) достаточно указать лишь одно из двух включений $A_1 \subseteq A_2$ или $B_2 \subseteq B_1$, поскольку в силу антимонотонности отображений «$'$» из одного из них всегда следует другое. Согласно (4), если $(A_1, B_1) \sqsubseteq (A_2, B_2)$, формальное понятие (A_2, B_2) можно считать более общим, чем понятие (A_1, B_1), поскольку оно имеет меньший набор характерных признаков, а значит, большее число объектов, обладающих этими признаками.

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

$$(A_1, B_1) \sqcap (A_2, B_2) = (A_1 \cap A_2, (A_1 \cap A_2)'), \quad (5)$$

$$(A_1, B_1) \sqcup (A_2, B_2) = ((B_1 \cap B_2)', B_1 \cap B_2). \quad (6)$$

Тогда частично упорядоченное множество $(FC_K, \sqsubseteq)$ образует решетку $L_K = (FC_K, \sqcap, \sqcup)$. Операции $\sqcup$ и $\sqcap$, установленные соотношениями (5) и (6), удовлетворяют всем необходимым для решеток законам ассоциативности, коммутативности, идемпотентности и поглощения [9]. Эта решетка называется решеткой формальных понятий контекста $K = (G, M, I)$. Известно, что L_K является полной решеткой [6]. Нулем решетки L_K является формальное понятие (M', M), содержащее все признаки контекста $K = (G, M, I)$, а единицей – формальное понятие (G, G'), в котором объем – множество всех объектов рассматриваемого контекста.

Решетка L_K связывает все элементы частично упорядоченного множества FC_K в определенную иерархическую структуру. Чем выше уровень расположения формального понятия в L_K, тем более общим по отношению к формальным понятиям, находящимся ниже в L_K, оно является. Таким образом, решетка L_K – это формализованное представление множества формальных понятий и связей между ними в смысле отношения (4). Причем каждое формальное понятие этой решетки определяет множество однородных объектов исследуемой предметной области со своим специфичным набором признаков. На основе решетки формальных понятий и методов АФП можно решать задачу бинарной классификации по прецедентам.

Задача бинарной классификации по прецедентам и алгоритм ее решения

Известны различные формулировки задачи классификации [3, 7]. Задача бинарной классификации по прецедентам традиционно формулируется следующим образом. Пусть задано конечное множество объектов G, разделенное на два класса G^+ и G^-, $G^+ \cap G^- = \varnothing$, $G^+ \cup G^- = G$. Такое разбиение определено с помощью некоторой обучающей выборки и целевого бинарного признака z. Элементы множеств G^+ и G^- называют положительными и отрицательными прецедентами соответственно. Все

объекты из G описаны через конечное множество признаков M, которое задается (0, 1)-матрицей T, кодирующей наличие или отсутствие признака $m \in M$ для объекта $g \in G$. Пусть задан некоторый объект $x \notin G$. Считается, что он обладает множеством признаков $M_x \subseteq M$. Требуется найти решающее правило, которое для объекта x определяет класс принадлежности. Решающее правило должно приводить к отказу от классификации, когда принадлежность объекта x к тому или иному классу не может быть однозначно определена.

Для описания данной задачи в терминах АФП достаточно лишь уточнить вид представления классов G^+ и G^-. С этой целью сопоставим классу G^+ положительный контекст $K^+ = (G^+, M, I^+)$, а классу G^- – отрицательный контекст $K^- = (G^-, M, I^-)$. Существование в I^+ пары (g, m) означает, что объект $g \in G^+$ имеет признак $m \in M$. Аналогично принадлежность пары (g, m) к I^- говорит о том, что объекту $g \in G^-$ присущ признак $m \in M$. Таким образом, бинарная матрица T разбивается на две подматрицы, соответствующие отношениям инцидентности I^+ и I^-.

Существуют различные алгоритмы классификации на основе АФП. К ним относятся алгоритмы Rulearner, GALOIS, GRAND, CITREC, CLNN & CLNB и LEGAL, использующие всю решетку понятий или ее некоторое подмножество [10], и алгоритмы, основанные на гипотезах [7]. В данной статье задача бинарной классификации по прецедентам решается с помощью гипотез.

Гипотезой называется некоторый набор признаков, который присутствует в описании объектов одного класса и не присутствует в описании объектов другого класса. Гипотезы извлекаются из решеток формальных понятий L_{K+} и L_{K-}, построенных для контекстов K^+ и K^- соответственно. Содержание B^+ формального понятия $(A^+, B^+) \in L_{K+}$ называется положительной гипотезой, если не существует такого формального понятия $(A^-, B^-) \in L_{K-}$, что $B^+ \subseteq B^-$. В противном случае множество признаков B^+ называется фальсифицированной положительной гипотезой. Аналогичным образом определяются отрицательные гипотезы и фальсифицированные отрицательные гипотезы: содержание B^- формального понятия $(A^-, B^-) \in L_{K-}$ считается отрицательной гипотезой, если не существует такого формального понятия $(A^+, B^+) \in L_{K+}$, что $B^- \subseteq B^+$, иначе B^- является фальсифицированной отрицательной гипотезой.

Решающее правило бинарной классификации по прецедентам для объекта x можно сформулировать следующим образом [7]:

– объект x относится к классу G^+, если множество M_x включает хотя бы одну положительную гипотезу и не включает ни одной отрицательной гипотезы; в противном случае объект x относится к классу G^-;

– отказ от классификации происходит, если M_x не включает в качестве подмножеств ни положительные, ни отрицательные гипотезы, или если M_x включает как положительные, так и отрицательные гипотезы.

Процесс решения задачи бинарной классификации на основе гипотез состоит из пяти этапов:

1-й этап – предобработка исходных контекстов;

2-й этап – нахождение формальных понятий в K^+ и K^-;

3-й этап – построение решеток L_{K+} и L_{K-};

4-й этап – выявление гипотез;

5-й этап – применение решающего правила бинарной классификации для объекта $x \notin G$.

На первом этапе производится предобработка исходных контекстов с целью уменьшения их размеров. Предобработка выполняется так, чтобы не изменились число и состав формальных понятий в L_{K+} и L_{K-}. Сокращение может затрагивать как множество объектов, так и множество признаков. Возможные случаи и алгоритмы их обработки рассмотрим применительно к положительному контексту и соответствующей ему матрице инцидентности.

Случай 1 *(дубликаты строк).*

Пусть в $K^+ = (G^+, M, I^+)$ существует множество объектов $A = \{g_1, g_2\}$, таких, что $g_1' = g_2' = B$. Тогда $A'' = (g_1' \cap g_2')' = (B \cap B)' = (B)' = A$, то есть A является замкнутым множеством. Следовательно, объект g_2 можно удалить из K^+ и не учитывать при вычислении положительных формальных понятий. При построении решетки L_{K+} объект g_2 необходимо добавить в объемы тех формальных понятий, в которые вошел объект g_1.

Случай 2 *(нулевые строки и столбцы).*

Если в $K^+ = (G^+, M, I^+)$ существует такой объект g, что $g' = \varnothing$, то $g'' = (g')' = (\varnothing)' = G^+$. Аналогично, если в $K^+ = (G^+, M, I^+)$ имеется признак $m \in M$, такой, что $m' = \varnothing$, то $m'' = (m')' = (\varnothing)' = M$. Поэтому на момент вычисления положительных формальных понятий объект g и признак m следует отбросить, а затем при построении L_{K+} объект g добавить в единицу, а признак m – в ноль этой решетки.

Случай 3 (*единичные строки и столбцы*).

Если в контексте $K^+ = (G^+, M, I^+)$ существует такой объект $g \in G^+$, что $g' = M$, то $g'' = (g')' = (M)' = g$. Поэтому объект g надо опустить при нахождении формальных понятий, но затем добавить в решетку L_{K+} новое формальное понятие (g, M), а объемы всех ранее полученных положительных формальных понятий пополнить объектом g. Аналогично, если имеется такой признак m, что $m' = G^+$, m вначале необходимо опустить и потом добавить в содержание всех формальных понятий решетки L_{K+}.

На втором и третьем этапах выявляются формальные понятия в исходных контекстах K^+ и K^-, прошедших предобработку. Простейшим способом осуществления этих действий является пере-

бор всех различных подмножеств множества признаков (их число, как правило, значительно меньше числа объектов) с вычислением для каждого из них замыкания по формуле (3). Затем на основе (4)–(6) строятся решетки L_{K+} и L_{K-} для контекстов K^+ и K^- соответственно.

На четвертом и пятом этапах выявляются положительные гипотезы, фальсифицированные положительные гипотезы, отрицательные гипотезы и фальсифицированные отрицательные гипотезы путем проверки отношения включения содержаний соответствующих формальных понятий. После этого в соответствии с приведенным выше решающим правилом классификации принимается решение о том, чтобы или отнести объект x к G^+ или к G^-, или констатировать отказ от классификации. Следует отметить, что на четвертом и пятом этапах процесса решения задачи бинарной классификации может быть использован не только алгоритм классификации на основе гипотез, но и любой другой алгоритм, базирующийся на АФП.

Проблема построения решетки формальных понятий и приемы снижения сложности вычислений

Рассмотренные выше задачи концептуального моделирования и бинарной классификации опираются на решетки формальных понятий. Известно, что задача порождения для заданного контекста всех формальных понятий и построения решетки формальных понятий является NP-трудной. Обоснование этого факта дано в [5]. Высокая вычислительная сложность данной задачи объясняется тем, что число формальных понятий может экспоненциально зависеть от размера контекста. Например, это имеет место для контекстов вида $K = (G, M, \neq)$. Поэтому время, необходимое на выявление формальных понятий в контексте $K = (G, M, I)$ и построение решетки, в худшем случае составляет $O(|FC_K| \cdot |G|^2 \cdot |M|)$, где $|FC_K|$ – число формальных понятий. Далее предлагаются два приема снижения вычислительной сложности этого процесса.

Прием 1: уменьшение размера величин $|G|$ и $|M|$ с помощью алгоритмов обработки случаев 1–3.

Эти случаи описаны выше, там же доказана корректность их применения. Время реализации приема 1 составляет $O(|G| \cdot |M|)$.

Прием 2: декомпозиция контекста – разделение контекста на полиномиальное число боксов (с последующим поиском формальных понятий в каждом из выделенных боксов).

Введем понятие бокса через объектные и признаковые формальные понятия контекста $K = (G, M, I)$. Назовем объектным понятием формальное понятие вида (g'', g'), где $g \in G$, а признаковым понятием – формальное понятие вида (m', m''), где $m \in M$. Таким образом, каждому объекту из G соответствует одно объектное понятие, и каждому признаку из M – одно признаковое понятие. Следовательно, для контекста $K = (G, M, I)$ число объектных понятий равно $|G|$, а число признаковых понятий составляет $|M|$. Заметим, что объектное понятие (g'', g') имеет самое большое по размеру содержание g' среди других формальных понятий, имеющих в объеме объект g, а признаковое понятие (m', m'') – самый большой объем m' среди других понятий, имеющих в содержании признак m. Это следует из антимонотонности соответствий Галуа, указанных в утверждениях 1 и 2.

Обозначим через $O_K = \{(g'', g') \mid \forall\, g \in G\} \subseteq FC_K$ множество всех объектных понятий и через $S_K = \{(m', m'') \mid \forall\, m \in M\} \subseteq FC_K$ множество всех признаковых понятий контекста $K = (G, M, I)$. Заметим, что множества O_K и S_K могут иметь непустое пересечение. Выберем два формальных понятия $(g'', g') \in O_K$ и $(m', m'') \in S_K$. Если для них верно отношение порядка $(g'', g') \sqsubseteq (m', m'')$ или, то же самое, выполняются условия

$$g'' \subseteq m' \text{ и } m'' \subseteq g', \qquad (7)$$

то пару (m', g') назовем боксом контекста $K = (G, M, I)$, образованным элементами $g \in G$ и $m \in M$. Пусть формальное понятие $(A, B) \in FC_K$ вложено в бокс (m', g') контекста $K = (G, M, I)$, если $A \subseteq m'$ и $B \subseteq g'$. Всякий бокс (m', g') не является пустым, поскольку согласно (7) в него всегда вложены формальные понятия $(g'', g') \in O_K$ и $(m', m'') \in S_K$.

Рассмотрим некоторый бокс (m', g'), образованный элементами $g \in G$ и $m \in M$ контекста $K = (G, M, I)$. Очевидно, что данный бокс определяет некоторую подматрицу матрицы T и образует подконтекст (G, M, C) контекста $K = (G, M, I)$, где $C \subseteq I$. При этом $(x, y) \in C$, если и только если $x \in m'$ и $y \in g'$. Соответствие между боксами и формальными понятиями контекста $K = (G, M, I)$ устанавливает утверждение 3 [11], подтверждающее корректность приема 2.

Утверждение 3. Для всякого контекста $K = (G, M, I)$ и любой пары множеств (A, B), где $\varnothing \neq A \subseteq G$, $\varnothing \neq B \subseteq M$, справедливы высказывания:

а) если (A, B) – формальное понятие контекста $K = (G, M, I)$, то всегда в этом контексте существует бокс (m', g'), образованный элементами $g \in G$ и $m \in M$, причем, возможно, не единственный, в который это формальное понятие вкладывается;

б) если (A, B) – формальное понятие подконтекста (G, M, C), соответствующего некоторому боксу (m', g') контекста $K = (G, M, I)$, то оно также является формальным понятием контекста $K = (G, M, I)$.

Если в контексте $K = (G, M, I)$ имеются формальные понятия $(G, \varnothing)$ и $(\varnothing, M)$, то для них невозможно установить признаковые и объектные понятия, поэтому они не вкладываются ни в один из боксов данного контекста. Их наличие необходимо просто учитывать при построении решеток.

Однократное формирование боксов для контекста $K = (G, M, I)$ включает в себя следующие действия: нахождение всех объектных и признаковых понятий; проверка условия (7) для каждой пары таких формальных понятий и формирование боксов. Число анализируемых пар, проверок и полученных боксов всегда не более чем $|I| = |G| \cdot |M|$. Поэтому время формирования боксов составляет $O(|I| \cdot (|G| + |M|))$. В худшем случае может быть найден только один бокс, совпадающий с исходным контекстом, и тогда декомпозиция контекста на боксы не дает эффекта. Это возможно, например, для контекста, полностью заполненного единицами. Однако реальные контексты, как правило, разлагаются на разумное число боксов. Важно отметить, что процесс разбиения контекста на боксы может быть организован итерационно, ведь каждый выявленный бокс может быть вновь разбит на боксы. Но если данный процесс продолжать до тех пор, пока все боксы выродятся в формальные понятия, это может привести к экспоненциальному числу боксов, а значит, и к экспоненциальному времени их построения. Для получения полиномиального числа боксов рекомендуется ограничиваться константным числом итераций.

Описание программы и результаты вычислительных экспериментов

Рассмотренные выше алгоритмы решения задач концептуального моделирования и бинарной классификации по прецедентам, а также приемы снижения вычислительной сложности этих алгоритмов реализованы в программе FCACorpus (язык программирования Delphi). Функция визуализации решеток формальных понятий не была включена в программу FCACorpus, так как ее применение целесообразно только для контекстов сравнительно небольших размеров. Визуализировать решетку можно всегда с помощью специальных программных средств [12].

Для оценки результативности приемов снижения сложности вычислений, реализованных в программе FCACorpus, были выполнены вычислительные эксперименты. Использовались контексты с числом объектов 15, 18, 20 и числом признаков 15. Эти контексты были сформированы на основе паспортов фольклорных произведений, взятых из Национального корпуса тувинского языка. Для каждого контекста $K = (G, M, I)$ осуществлялся поиск множества FC_K всех формальных понятий без разбиения и с однократным разбиением этого контекста на боксы. Результаты вычислительных экспериментов приведены в таблице 1, где $|G|$ – количество объектов контекста; $|FC_K|$ – число найденных формальных понятий; N – количество образованных боксов; t – время выполнения программы. Вычислительные эксперименты выполнялись на компьютере с процессором Intel® Core™ i7-720QM Processor (6M Cache, 1.60 GHz) и ОЗУ размером 4 ГБ.

Таблица 1

Результаты экспериментов

Table 1

The experimental results

Вычисление всех формальных понятий контекста	$\|G\|$	$\|FC_K\|$	N	t, мс
Без разбиения на боксы	15	36	–	480
С разбиением на боксы		36	12	66
Без разбиения на боксы	18	73	–	12480
С разбиением на боксы		73	23	120
Без разбиения на боксы	20	98	–	30519
С разбиением на боксы		98	40	150

Как видно из таблицы 1, количество и состав полученных формальных понятий в обоих случаях (без разбиения на боксы, с разбиением на боксы) полностью совпадают. Однако применение боксов дает значительный выигрыш во времени – время выполнения программы в этом случае уменьшается в 10–20 раз. Эксперименты на случайно сгенерированных контекстах различной размерности, показали, что, чем больше объектов и признаков содержит анализируемый контекст, тем больше выигрыш во времени.

Результаты концептуального моделирования и классификации произведений тувинского фольклора

Рассмотрим пример применения АФП и разработанных программ для филологических исследований, направленных на концептуальное моделирование произведений тувинского фольклора и определение их принадлежности к жанру героического эпоса. Для решения этих задач необходимо сформировать соответствующие контексты на основе паспортов произведений. Паспорт произведения – это набор признаков, характеризующих семантические, синтаксические и морфологические особенности этого произведения. Например, паспорт произведения тувинского фольклора содержит информацию о сказителе, сведения о пространственно-временном периоде написания, жанровые и сюжетные особенности произведения. Всего выделено 14 признаков, приведенных в таблице 2.

В таблице 3 представлен бинарный контекст $K^+ = (G^+, M, I^+)$ для четырех фольклорных произведений, где G^+ = {«Арзылаң-Кара аттыг Хунан-Кара», «Мөрүн-Хүлүк», «Өлээдей-Мерген», «Элестей ашак»}; $M = \{s_1, s_2, s_3, s_4, a_1, a_2, q_1, q_2, c_1, c_2, c_3, t_1, t_2, t_3\}$; I^+ – отношение инцидентности между G^+ и M.

Известно, что все произведения из G^+ относятся к жанру тувинского героического эпоса [13]. В таблице 3 названия произведений заменены их поряд-

ковыми номерами. Единичный (нулевой или пустой) элемент этой таблицы указывает на то, что соответствующее литературное произведение обладает (не обладает) тем или иным признаком.

Таблица 2

Паспорт произведения тувинского фольклора

Table 2

Passport of a Tuvan folklore work

Идентификатор признака	Значение признака
s_1	Сказитель Кашкак
s_2	Сказитель Хертек
s_3	Сказитель Ооржак
s_4	Другой сказитель или народ
a_1	Горный ареал
a_2	Степной ареал
q_1	Есть богатырь
q_2	Нет богатыря
c_1	Сюжет «Сватовство»
c_2	Сюжет «Сестра добывает брату суженую»
c_3	Другой сюжет
t_1	Зачин «Эрте шагның эктинде, бурун шагның мурнунда»
t_2	Зачин «Шыянам, эрте бурунгу шагда»
t_3	Зачин «Шыянам, эртенгиниң эртезинде бурунгунуң мурнунда»

Таблица 3

Контекст K^+ произведений тувинского героического эпоса

Table 3

The context K^+ of Tuvan heroic epic works

№	s_1	s_2	s_3	s_4	a_1	a_2	q_1	q_2	c_1	c_2	c_3	t_1	t_2	t_3
1			1		1		1		1			1		
2	1					1	1		1				1	
3		1			1		1			1				1
4			1		1		1			1		1		

Заметим, что контекст K^+ допускает предобработку согласно описанным выше случаям 2 и 3. Всего контекст K^+ порождает 10 формальных понятий, которые образуют решетку L_{K+} (рис. 1). Единицей этой решетки является формальное понятие $(G^+, \{q_1\})$, а нулем – $(\varnothing, M)$.

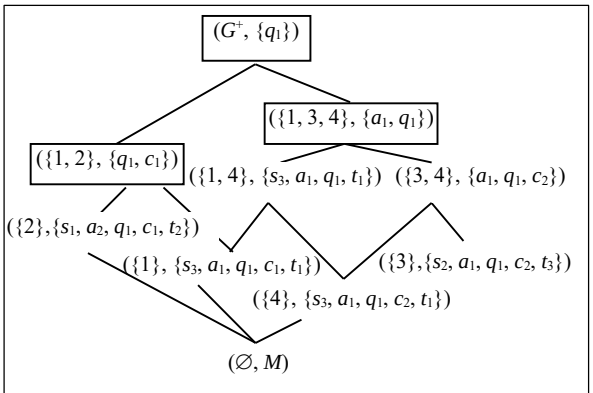

Рис. 1. Решетка формальных понятий L_{K+}

Fig. 1. The formal L_{K+} concept lattice

Решетка L_{K+} – концептуальная модель для множества произведений тувинского героического эпоса G^+, построенная и представленная в терминах АФП. Выявленные формальные понятия и семантические связи между ними позволяют сделать следующие выводы.

1. Для произведений, относящихся к жанру тувинского героического эпоса, характерно прежде всего наличие богатыря, так как единицей решетки L_{K+} является формальное понятие $(G^+, \{q_1\})$. Это понятие – самое общее по отношению ко всем другим формальным понятиям этой решетки. Ведь по определению решетки, чем выше уровень расположения формального понятия в L_{K+}, тем более общим по отношению к формальным понятиям, находящимся ниже в L_{K+}, оно является.

2. Произведениям тувинского героического эпоса присущи признаки Горный ареал или Сюжет «Сватовство». Эти признаки входят в содержание формальных понятий ($\{1, 3, 4\}$, $\{a_1, q_1\}$), ($\{1, 2\}$, $\{q_1, c_1\}$), расположенных в решетке L_{K+} уровнем ниже, чем формальное понятие $(G^+, \{q_1\})$, и выше всех других понятий.

3. Согласно формальному понятию ($\{1, 4\}$, $\{s_3, a_1, q_1, t_1\}$), для произведений героического эпоса, сказителем которых является Ооржак, специфичным является зачин «Эрте шагның эктинде, бурун шагның мурнунда».

Экспертами установлено, что указанные выводы соответствуют действительности, то есть являются филологически правильными. Каждый из указанных выводов – определенные знания о произведениях тувинского героического эпоса, представленных в K^+. Очевидно, что увеличение числа произведений в K^+ углубляет эти знания.

Рассмотрим теперь задачу бинарной классификации по прецедентам. Для этого сформируем отрицательный контекст $K^- = (G^-, M, I^-)$, состоящий из трех литературных произведений, которые не относятся к жанру тувинского героического эпоса (табл. 4). Здесь G^- = {«Чечен-Маанай и Тенек-Тулун», «Караты-Хаан биле Алдын-кыс», «Кыс-Халыыр»} или с использованием порядковых номеров произведений $G^- = \{5, 6, 7\}$. Следует отметить, что контекст K^- также допускает предобработку. Контекст K^- порождает 7 формальных понятий, которые образуют решетку L_{K-} (рис. 2). Единицей этой решетки является формальное понятие $(G^-, \{s_4\})$, а нулем – $(\varnothing, M)$.

Контексты $K^+ = (G^+, M, I^+)$, $K^- = (G^-, M, I^-)$ соответствуют двум классам G^+ и G^- произведений, разделенных по целевому бинарному признаку z = «произведение относится (не относится) к жанру героического эпоса». Пусть задано новое произведение x с множеством признаков $M_x = \{s_2, a_1, q_1, c_2, t_1\}$. Требуется для x определить класс, к которому его можно отнести.

Таблица 4

Контекст K^- произведений тувинского фольклора

Table 4

The context K^- of Tuvan folklore works

№	s_1	s_2	s_3	s_4	a_1	a_2	q_1	q_2	c_1	c_2	c_3	t_1	t_2	t_3
5				1		1		1			1		1	
6				1	1		1		1				1	
7				1		1		1			1			1

В решетке L_{K+} множества признаков $\{s_3, a_1, q_1, t_1\}$, $\{a_1, q_1, c_2\}$, $\{s_1, a_2, q_1, c_1, t_2\}$, $\{s_3, a_1, q_1, c_1, t_1\}$, $\{s_2, a_1, q_1, c_2, t_3\}$, $\{s_3, a_1, q_1, c_2, t_1\}$ являются положительными гипотезами, а $\{q_1\}$, $\{a_1, q_1\}$, $\{q_1, c_1\}$ – фальсифицированными положительными гипотезами. В решетке L_{K-} множества признаков $\{s_4\}$, $\{s_4, t_2\}$, $\{s_4, a_2, q_2, c_3\}$, $\{s_4, a_2, q_2, c_3, t_3\}$, $\{s_4, a_1, q_1, c_1, t_2\}$, $\{s_4, a_2, q_2, c_3, t_2\}$ определяют отрицательные гипотезы.

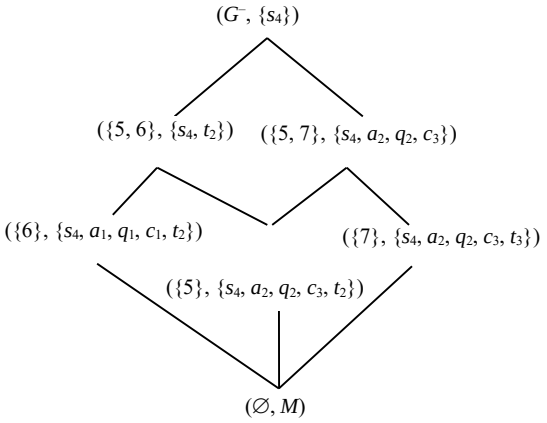

Рис. 2. Решетка формальных понятий L_{K^-}

Fig. 2. The formal L_{K-}concept lattice

По правилу бинарной классификации произведение x с набором признаков $M_x = \{s_2, a_1, q_1, c_2, t_1\}$ будет отнесено к классу G^+, то есть к жанру героического эпоса, так как M_x включает положительную гипотезу $\{a_1, q_1, c_2\}$ и не содержит отрицательных гипотез. Если $M_x = \{s_4, t_2\}$, то произведение x будет отнесено к классу G^-. При $M_x = \{q_1\}$ произойдет отказ от классификации. Применение метода скользящего контроля к используемому алгоритму классификации показало его вполне удовлетворительное качество [14].

Программа FCACorpus в настоящее время успешно используется в научно-образовательном центре «Тюркология» Тувинского государственного университета для филологических и лингвистических исследований естественно-языковых текстов, представленных в Национальном корпусе тувинского языка. В дальнейшем предполагается расширить функциональные возможности программы FCACorpus с целью повышения эффективности используемых в ней алгоритмов.

Литература

1. Салчак А.Я., Байыр-оол А.В. Электронный корпус тувинского языка: состояние, проблемы // Мир науки, культуры, образования. 2013. № 6. С. 408–409.

2. Бавуу-Сюрюн М.В. Вопросы создания электронных ресурсов тувинского языка: некоторые итоги, неотложные задачи и перспективы // Новые исследования Тувы. 2016, № 4. URL: http://nit.tuva.asia/nit/article/view/610 (дата обращения: 14.06.2017).

3. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. СПб: БХВ-Петербург, 2008. 384 с.

4. Богатырев М.Ю., Нуриахметов В.Р., Вакурин В.С. Методы анализа формальных понятий в информационных системах технической поддержки // Изв. ТулГУ: Технич. науки. 2013. Вып. 2. С. 25–36.

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

6. Ganter B., Wille R. Formal concept analyses: mathematical foundations. Springer Science and Business Media, 2012, 284 p.

7. Гуров С.И., Онищенко А.А. Классификация на основе АФП и бикластеризации: возможности подхода // Прикладная математика и информатика: тр. факульт. ВМК МГУ. 2011. Т. 38. С. 77–87.

8. Vlasov D.V. The methods of forming the theoretical concepts. Jour. of the Buryat State Univ., 2009, no. 6, pp. 37–41.

9. Биргоф Г. Теории решеток. М.: Наука, 1984. 568 с.

10. Meddouri N., Meddouri M. Classification methods based on formal concept analysis. CLA 2008, pp. 9–16.

11. Bykova V.V., Mongush Ch.M. On Algebraic Approach of R. Wille and B. Ganter in the Investigation of Texts. Jour. of Siberian Federal Univ.: Math. and Physics. 2017, no. 3, pp. 372–384.

12. Евтушенко С.А. Система анализа данных CONCEPT EXPLORER // КИИ-2000: сб. тр. VII Национальн. конф. по искусств. интеллекту М.: Физматлит, 2000. С. 127–134.

13. Орус-оол С.М. Тувинские героические сказания (текстология, поэтика, стиль). М.: МАКС Пресс, 2001. 422 с.

14. Воронцов К.В. Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики. М.: Физматлит, 2004. Т. 13. С. 5–36.

Software & Systems
DOI: 10.15827/0236-235X.119.487-495

Received 15.06.17
2017, vol. 30, no. 3, pp. 487–495

ALGORITHMS OF CONCEPTUAL MODELING AND TEXT CLASSIFICATION IN THE TUVAN LANGUAGE CORPUS

V.V. Bykova [1], *Dr. Sc. (Physics and Mathematics), Professor, bykvalen@mail.ru*
Ch.M. Mongush [1, 2], *Postgraduate Student, Lecturer, mongushchod91@yandex.ru*

[1] *Siberian Federal University, Svobodny Ave. 79, Krasnoyarsk, 660041, Russian Federation*
[2] *Tuvan State University, Lenina St. 36, Kyzyl, 667000, Russian Federation*

Abstract. The corpus is an information-linguistic system based on the collection of digitized texts in some language. Nowadays, the corpus of Tuvan language includes official and business documents and Tuvan literary works.

Expanding of the Tuvan corpus and deepening of the text processing level are continuening. These works lead to the tasks of a natural language text analysis. The main tasks is classification by precedents and conceptual modeling.

In order to solve these problems, the paper uses an algebraic approach, which is called the analysis of formal concepts. The paper proposes algorithms and programs for constructing a conceptual model of literary works collection and solving the problem of a binary classification by precedents. There are methods of reducing computational complexity of the considered algorithms.

The paper presents the results of computational experiments, which confirm the effectiveness of the proposed methods for reducing computation complexity. Finally, there are the results of conceptual modeling and binary classification of Tuvan folklore works.

Keywords: corpus, formal concept analysis, conceptual models of texts, classification algorithm, algorithms of reducing a context dimension.

References

1. Salchak A.Ya., Bayyr-ool A.V. Electronic housing of tuvan language: condition, issues. *Mir nauki, kultury, obrazovaniya* [The World of Science, Culture and Education]. 2013, no. 6, pp. 408–409 (in Russ.).

2. Bavuu-Syuryun M.V. Creating electronic resources on tuvan language: preliminary results, current challenges and prospects. *Novye issledovaniya Tuvy* [The New Research of Tuva]. 2016, no. 4. Available at: http://nit.tuva.asia/nit/article/view/610 (accessed June 14, 2017).

3. Barsegyan A.A., Kupriyanov M.S., Stepanenko V.V., Kholod I.I. *Tekhnologii analiza dannykh: Data Mining, Visual Mining, Text Mining, OLAP* [Data Analysis Technologies: Data Mining, Visual Mining, Text Mining, OLAP]. St. Petersburg, BHV-Peterburg Publ., 2008, 384 p.

4. Bogatyrev M.Yu., Nuriakhmetov V.R., Vakurin V.S. Methods of formal notion analysis in technical support information systems. *Izvestiya TulGU. Tekhnicheskie nauki* [News of the Tula State University. Technical Sciences]. 2013, iss. 2, pp. 25–36 (in Russ.).

5. Kuznetsov S.O. Machine Learning on the Basis of Formal Concept Analysis. *Avtomatika i telemekhanika* [Automation and Remote Control]. 2001, vol. 62, iss. 10, pp. 1543–1564.

6. Ganter B., Wille R. *Formal concept analyses: mathematical foundations*. Springer Science and Business Media Publ., 2012, 284 p.

7. Gurov S.I., Onishchenko A.A. Classification based on Formal Concept Analysis and biclasterization: the opportunities of the approach. *Prikladnaya matematika i informatika: Tr. fakulteta VMK MGU* [Applied Mathematics and Computer Science: Faculty of Computational Mathematics and Cybernetics of the Lomonosov MSU]. 2011, vol. 38, pp. 77–87.

8. Vlasov D.V. The methods of forming the theoretical concepts. *Jour. of the Buryat State Univ*. 2009, no. 6, pp. 37–41.

9. Birgof G. *Teorii reshetok* [Category Lattice]. Moscow, Nauka Publ., 1984, 568 p.

10. Meddouri N., Meddouri M. Classification Methods Based on Formal Concept Analysis. *CLA 2008*, pp. 9–16.

11. Bykova V.V., Mongush Ch.M. On Algebraic Approach of R. Wille and B. Ganter in the Investigation of Texts. *Jour. of Siberian Federal Univ.: Math. and Physics*. 2017, no. 3, pp. 372–384.

12. Evtushenko S.A. CONCEPT EXPLORER data analisys sistem. *Proc. 7th National Conf. on Artificial Intelligence KII-2000*. Moscow, Fizmatlit Publ., 2000, pp. 127–134 (in Russ.).

13. Orus-ool S.M. *Tuvinskie geroicheskie skazaniya (tekstologiya, poetika, stil)* [Tuvan Heroic Folk Tales (Textology, Poetics, Style)]. Moscow, Maks Press, 2001, 422 p.

14. Vorontsov K.V. A combinatorial approach to qualitative assessment of learning algorithms. *Matematicheskie voprosy kibernetiki* [Mathematical Problems of Cybernetics]. Moscow, Fizmatlit Publ., 2004, vol. 13, pp. 5–36 (in Russ.).

Примеры библиографического описания статьи

1. Быкова В.В., Монгуш Ч.М. Алгоритмы концептуального моделирования и классификации текстов в корпусе тувинского языка // Программные продукты и системы. 2017. Т. 30. № 3. С. 487–495. DOI: 10.15827/0236-235X.119.487-495.

2. Bykova V.V., Mongush Ch.M. Algorithms of conceptual modeling and text classification in the tuvan language corpus. *Programmnye produkty i sistemy* [Software & Systems]. 2017, vol. 30, no. 3, pp. 487–495 (in Russ.). DOI: 10.15827/0236-235X.119.487-495.