УДК 004.82
DOI 10.25205/1818-7900-2018-16-4-140-152

А. Б. Нугуманова [1], **Е. М. Байбурин** [1], **М. Е. Мансурова** [2], **В. Б. Барахнин** [3, 4]

[1] *Восточно-Казахстанский государственный университет им. С. Аманжолова*
ул. 30-й Гвардейской Дивизии, 34, Усть-Каменогорск, 070002, Казахстан

[2] *Казахский национальный университет им. аль-Фараби*
пр. аль-Фараби, 71, 050040, Алматы, Казахстан

[3] *Институт вычислительных технологий СО РАН*
пр. Акад. Лаврентьева, 6, Новосибирск, 630090, Россия

[4] *Новосибирский государственный университет*
ул. Пирогова, 1, Новосибирск, 630090, Россия

{anugumanova, ebaiburin}@vkgu.kz, mansurova.madina@gmail.com, bar@ict.nsc.ru

АВТОМАТИЧЕСКОЕ ИЗВЛЕЧЕНИЕ РЕШЕТОК ПОНЯТИЙ ИЗ МЕДИЦИНСКИХ ТЕКСТОВ НА ОСНОВЕ КОМБИНАЦИИ АНАЛИЗА ФОРМАЛЬНЫХ ПОНЯТИЙ И ТЕХНОЛОГИЙ БУТСТРАППИНГА [*]

Рассматривается новый способ извлечения понятий из текстов предметной области на основе комбинации анализа формальных понятий и бутстрап-технологии информационного поиска. Анализ формальных понятий представляет собой мощный аппарат автоматического вывода понятий предметной области, однако он рассчитан на высокое качество входных данных, без пропусков и неточностей. Получение таких наборов данных напрямую из текстов затруднено в силу сильной разреженности текстовых корпусов. Соответственно, представляется перспективным улучшение качества входных данных за счет применения бутстраппинга – технологии, обеспечивающей интеллектуальный поиск фрагментированной информации в сети Интернет. Цель данной работы – показать, что при правильном выборе исходных шаблонов поиска бутстраппинг, основанный на использовании открытых ресурсов Интернета как ценных источников знаний, превращается в эффективный инструмент поддержки концептуального моделирования.

Ключевые слова: анализ формальных понятий, бутстраппинг, извлечение информации, поверхностный лингвистический анализ, информационный поиск.

Введение

В данной работе мы предлагаем новый способ автоматического извлечения понятий из текстов медицинской тематики, основанный на заполнении пропусков в сильно разреженных матрицах совместной встречаемости терминов. Наш подход опирается на строгие и изящные формулировки анализа формальных понятий – метода, который, как отмечается в работе [1], использует язык алгебры для описания понятий и их иерархий.

Входные данные для анализа формальных понятий представляются в виде объектно-признаковой таблицы, отражающей распределение признаков по объектам предметной об-

[*] Работа выполнена при частичной поддержке РФФИ (проект 18-07-01457).

Нугуманова А. Б., *Байбурин Е. М.*, *Мансурова М. Е.*, *Барахнин В. Б.* Автоматическое извлечение решеток понятий из медицинских текстов на основе комбинации анализа формальных понятий и технологий бутстраппинга // Вестн. НГУ. Серия: Информационные технологии. 2018. Т. 16, № 4. С. 140–152.

ласти. Таблица является бинарной: если объект, указанный в строке таблицы, обладает признаком, указанном в столбце таблицы, то на пересечении соответствующих строки и столбца ставится 1, иначе 0. Математический аппарат анализа формальных понятий позволяет выделить в этой таблице множества объектов, обладающих одинаковыми наборами признаков. Считается, что каждое такое множество объектов и их признаков образует одно формальное понятие. Объекты этого множества называются объемом понятия (экстенсионалом), а признаки этого множества – содержанием понятия (интенсионалом).

Особенность нашей работы заключается в том, что в качестве входной объектно-признаковой таблицы мы используем матрицу совместной встречаемости терминов. Термины, образующие строки и столбцы матрицы, мы получаем, извлекая из текстов медицинской тематики конструкции, удовлетворяющие лексико-синтаксическим шаблонам вида «Существительное + Существительное». Один из самых результативных шаблонов этого вида представляет собой пару «Существительное + Существительное в родительном падеже». Так, например, при помощи этого шаблона мы извлекаем такие словосочетания, как «приступ холецистита», «приступ панкреатита», «воспаление нерва», «нарушение кровоснабжения», «обострение гайморита», «осложнение гайморита» и т. д. Термины, стоящие в указанных парах на первом месте, мы записываем в столбцы матрицы (признаки), а термины, стоящие на втором месте, – в строки (объекты). Также возможна работа и с транспонированной матрицей, тогда термины, стоящие на первом месте, мы записываем в строки, а термины, стоящие на втором месте, – в столбцы.

К сожалению, полученную таким способом входную матрицу невозможно напрямую использовать для извлечения формальных понятий, так как для этого она слишком неоднородная и разреженная. Дело в том, что строки этой матрицы отражают весь гетерогенный набор объектов, встречающихся в текстах медицинской тематики: названия заболеваний, лекарств, методов лечения, органов человеческого тела и т. д. Соответственно и столбцы этой матрицы отражают весь спектр разнородных признаков, каждый из которых присущ только одной «своей» группе объектов (например, признаками «обострение» и «осложнение» описываются заболевания, но не органы, а признаками «воспаление» и «отек» описываются органы, но не заболевания). Как следствие, увеличивается и без того сильная разреженность, изначально присущая текстовому корпусу.

Для устранения неоднородности мы подвергаем исходную матрицу кластеризации, в результате чего из нее получаются более однородные и менее разреженные матрицы меньшего размера – кластеры, представляющие собой группы объектов со сходными признаками. Эти матрицы мы используем как отправные шаблоны для бутстраппинга – технологии, позволяющей находить недостающую информацию по ее начальным фрагментам. В частности, мы извлекаем из указанных матриц все нулевые пары «объект – признак» и для каждой такой пары формируем поисковый запрос к Интернету как к более репрезентативному корпусу текстов. Таким образом, для каждой нулевой пары «объект – признак» мы проверяем, существует ли устойчивый контекст употребления данного объекта с данным признаком в таком сверхбольшом текстовом корпусе, как Интернет. Если такой контекст существует и является устойчивым, то мы заменяем соответствующий нулевой элемент в матрице на единичный.

Только после восстановления пропусков и увеличения количества значимой информации в матрицах мы используем их для извлечения понятий. Как отмечалось выше, для этого мы применяем аппарат анализа формальных понятий, который позволяет не только автоматически выводить понятия предметной области, но и формировать их иерархии. Цель нашей работы заключается в том, чтобы продемонстрировать на конкретном корпусе текстов, как работает предлагаемый подход и насколько эффективно он решает задачу автоматического извлечения понятий.

В соответствии с поставленной целью дальнейшее изложение работы ведется следующим образом. Сначала мы приводим основные положения анализа формальных понятий. В последующих разделах мы описываем 4 основных этапа предлагаемого подхода.

- Этап 1: построение исходной объектно-признаковой матрицы на основе поверхностного лингвистического анализа.
- Этап 2: разбиение исходной объектно-признаковой матрицы на ряд однородных матриц посредством кластеризации.

- Этап 3: восстановление пропусков в полученных однородных матрицах методом бутстраппинга.
- Этап 4: извлечение из восстановленных матриц понятий предметной области и построение их иерархий на основе анализа формальных понятий.

В заключение мы излагаем основные выводы и приводим план будущей работы.

Математический аппарат анализа формальных понятий

Входными данными для анализа формальных понятий (АФП) служит информация о распределении признаков среди объектов предметной области [2]. Указанная информация записывается в виде так называемого формального контекста *K*, который представляет собой тройку

$$K = \langle G, M, I \rangle,$$

где *G* – это множество объектов, *M* – множество признаков, *I* – соответствие между *G* и *M*: *gIm* означает, что объект $g \in G$ обладает признаком $m \in M$. Оставим пока в стороне вопрос, каким образом из предметной области выбираются объекты и их признаки для формирования формального контекста. Предположим, что формальный контекст уже задан, и покажем, как в заданном контексте выделяются формальные понятия.

Пусть в формальном контексте $K = \langle G, M, I \rangle$ выбраны произвольные подмножества объектов $A \subseteq G$ и признаков $B \subseteq M$. Операторы Галуа для указанных подмножеств определяются следующим образом:

- $A' = \{ m \in M \mid \forall g \in A \; gIm \}$, т. е. A' – это множество признаков, которыми обладают все объекты из *A*;
- $B' = \{ g \in G \mid \forall m \in B \; gIm \}$, т. е. B' – это множество признаков, которыми обладают все объекты из *B*.

Тогда формальным понятием контекста *K* называется пара вида (A, B), $A \subseteq G$ и $B \subseteq M$, такая что $A' = B$ и $B' = A$. Множества *A* и *B* называются соответственно объемом и содержанием формального понятия (A, B).

Между содержанием и объемом понятия существует обратная зависимость: чем больше содержание, тем меньше объем. Другими словами, чем больше признаков содержит данное понятие, тем меньше объектов оно охватывает, и наоборот.

Два понятия (A_1, B_1) и (A_2, B_2) называют частично упорядоченными: $(A_1, B_1) \le (A_2, B_2)$, если объем первого понятия входит (вложен) в объем второго: $A_1 \subseteq A_2$. Множество всех понятий контекста *K*, упорядоченных по вложению их объемов, называется решеткой понятий. Для визуального представления решеток применяются диаграммы Хассе, где сверху показаны наименьшие по объему понятия, а снизу – наибольшие [1; 3].

Поясним основные положения АФП на конкретном примере.

На рис. 1 представлен фрагмент формального контекста, сформированного на основе автоматической обработки корпуса медицинских текстов. Контекст задан обычным для АФП способом – в виде объектно-признаковой таблицы.

A	B	C	D	E	F	G	H	I	J	K	L
	"наличие"	"уровень"	"количеств...	"содержан...	"активност...	"раствор"	"концентр...	"потребле...	"накоплен...	"добавлен...	"утилизац...
"амилаза"		×	×	×	×	×	×			×	
"аминотра...		×	×	×	×		×				
"глюкоза"	×	×	×	×	×	×	×	×	×	×	×
"креатинин"	×	×	×	×	×		×		×		×
"лактат"	×	×	×	×	×	×	×	×	×	×	×
"мочевина"	×	×	×	×	×		×		×		×
"трансфер...		×	×	×	×		×				

Рис. 1. Фрагмент формального контекста

В табл. 1 перечислены все четыре формальных понятия, выведенных из рассматриваемого формального контекста. Первое формальное понятие самое широкое по объему, оно включает в себя все 7 объектов, поскольку все они описываются такими общими признаками, как «уровень», «количество», «содержание», «активность» и «концентрация». Утрируя, можно сказать, что в реальности этому формальному понятию соответствует понятие вещества. Второе формальное понятие включает в себя 3 объекта, которые обладают такими дополнительными признаками, как «раствор» и «добавление». Аналогично можно сказать, что в реальности этому формальному понятию соответствует понятие растворимого вещества. Самым узким по объему формальным понятием является четвертое, оно содержит всего 2 объекта: «глюкоза» и «лактат». Этому формальному понятию соответствует понятие вещества, которое можно употреблять в пищу.

Таблица 1

Формальные понятия, выведенные из формального контекста на рис. 1

№	Понятие	Объем понятия	Содержание понятия
1	(A_1, B_1)	A_1 ={амилаза, аминотрансфераза, глюкоза, креатинин, лактат, мочевина, трансфераза}	B_1 ={уровень, количество, содержание, активность, концентрация}
2	(A_2, B_2)	A_2 ={амилаза, глюкоза, лактат}	$B_2 = B_1 \cup$ {раствор, добавление}
3	(A_3, B_3)	A_3 ={глюкоза, креатинин, лактат, мочевина}	$B_3 = B_1 \cup$ {утилизация, накопление, наличие}
4	(A_4, B_4)	A_4 ={глюкоза, лактат}	$B_4 = B_2 \cup B_3 \cup$ {потребление}

На рис. 2 изображена диаграмма Хассе, состоящая из двух решеток понятий, построенных на основе рассматриваемого формального контекста. Однако, анализируя полученные иерархии и классы объектов, эксперт предметной области мог бы заметить, что они не корректны. Например, креатинин и мочевина несправедливо исключены из понятия «растворимое вещество», – они тоже могут растворяться. В этом и состоит ключевая проблема анализа формальных понятий: если контекст является разреженным и ограниченным, то выходные понятия и их иерархии будут формироваться некорректно.

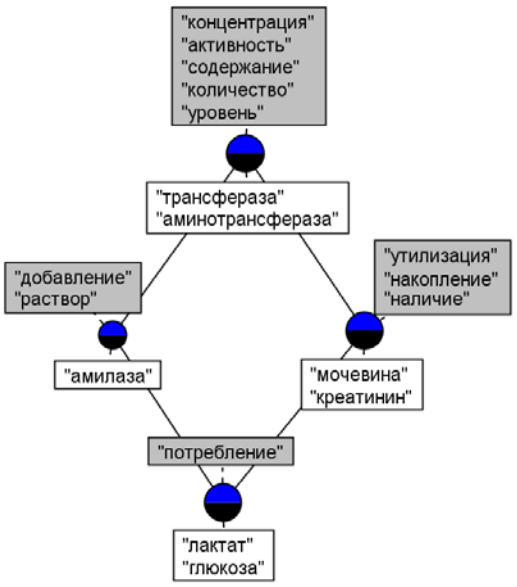


Рис. 2. Диаграмма Хассе

Построение входной объектно-признаковой матрицы

Для автоматического построения входной объектно-признаковой матрицы на основе корпуса текстов мы используем поверхностный лингвистический анализ (shallow linguistic analysis). Поверхностный лингвистический анализ – это весьма популярный подход к обработке естественного языка, когда речь идет о создании практических приложений в области анализа текстов, ориентированных на конкретные интересы пользователей [4].

Как следует из его названия, подход не фокусируется на глубоком (многоуровневом и многоаспектном) разборе текстов, его цель – быстрое выделение и анализ только тех текстовых фрагментов, которые содержат релевантную с точки зрения пользователя информацию [5]. Благодаря этому подход демонстрирует

высокую производительность и устойчивость на сколь угодно больших корпусах текстов. Вместе с тем он требует обязательной спецификации того, что считать релевантной информацией, поскольку для выделения такой информации из текста необходима настройка шаблонов, определяющих соответствие между естественно-языковыми высказываниями и знаниями предметной области.

В этом и заключается основной недостаток поверхностных лингвистических методов. Во-первых, как отмечается в [5], подобная «тонкая» настройка методов под конкретную предметную область препятствует их переносу на другие предметные области или задачи. Во-вторых, даже внутри исходной предметной области остается неохваченным большой пласт релевантной информации, так как живая структура естественного языка плохо поддается шаблонизации.

Применительно к нашей задаче релевантной является информация о медицинских объектах и их признаках. Под медицинскими объектами мы понимаем объекты и явления, которые описываются в текстах медицинской тематики: это заболевания, диагнозы, способы лечения, лекарственные препараты, тело человека и т. д.

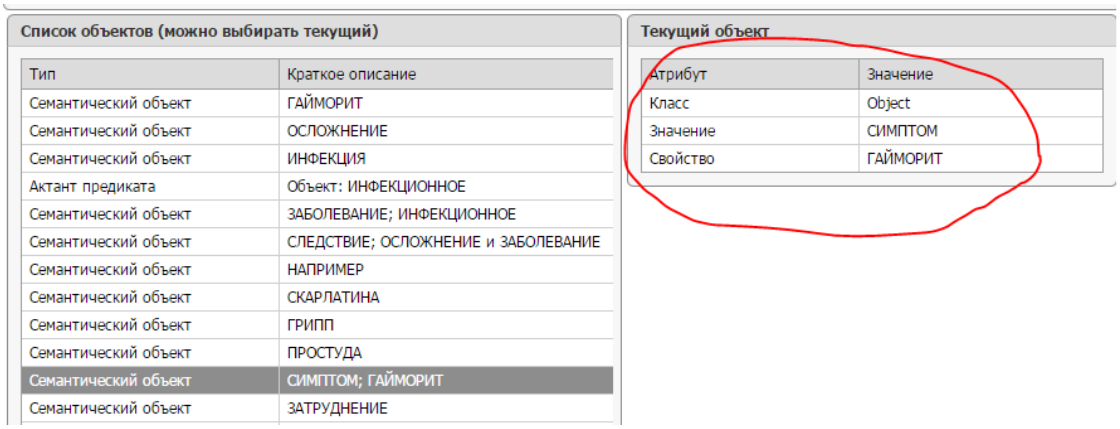
Новизна нашего подхода заключается в том, что указанные объекты и их признаки мы ищем в текстах корпуса на основе поверхностных лингвистических шаблонов вида «Существительное + Существительное в родительном падеже». Другими словами, мы формируем некоторое начальное множество объектов и их признаков путем извлечения из текстов пар слов, удовлетворяющих заданному шаблону. В качестве признаков мы рассматриваем слова, стоящие в парах на первом месте, а в качестве объектов – слова, стоящие на втором месте.

Однако предлагаемый нами способ поиска объектов и признаков наследует недостатки, присущие поверхностным лингвистическим методам. Во-первых, он плохо адаптируется к изменениям параметров исходной задачи. Во-вторых, если использовать предлагаемый шаблон напрямую, без дополнительного разбора предложения, то он не покрывает всего множества релевантных объектов и их признаков. Например, объекты и их признаки, разделенные определениями, будут потеряны, т. е. конструкции вида «воспаление нерва», «осложнение отита» и т. п. будет извлечены, в то время как конструкции вида «воспаление зрительного нерва», «осложнение гнойного отита» и т. п. будут пропущены. Поэтому нам приходится расширять данный шаблон.

Для практической реализации предлагаемого способа извлечения объектов и их свойств мы используем программный пакет PullEnti – свободно распространяемый набор средств разработки (SDK) для создания приложений по анализу текстов [6; 7]. Мы подключаем указанные средства разработки в наш проект, реализуемый на языке C#, как внешнюю DLL-библиотеку. На рис. 3 представлен пример работы SDK PullEnti по извлечению объектов и их признаков из текстового фрагмента, описывающего симптомы гайморита: «Обычно гайморит – это следствие осложнений после инфекционного заболевания, например, скарлатины, гриппа, простуды. Основные симптомы гайморита – затрудненное дыхание, постоянный насморк, заложенность носа и головная боль». Из данного фрагмента были извлечены такие пары объектов и их признаков, как «симптом гайморита», «заложенность носа», «следствие осложнения», «следствие заболевания».

На рис. 4 представлен еще один пример работы SDK PullEnti. Здесь объекты и их признаки извлекаются из текстового фрагмента, описывающего осложнение гнойной ангины: «Синдром Лемьера – редкое, но серьезное осложнение гнойной ангины, иногда ее называют постангинальным сепсисом». Из данного фрагмента были извлечены пары «синдром Лемьера» и «осложнение ангины». Как мы видим, несмотря на то что в этом фрагменте объекту «ангина» предшествует определение «гнойная», а признаку «осложнение» предшествует определение «серьезное», это не мешает программе правильно установить объект и его признак.

Извлеченные таким образом пары «объект – признак» служат исходным материалом для заполнения объектно-признаковой матрицы: все названия объектов (без повторений) записываются в строках матрицы, а все названия признаков (тоже без повторений) записываются в столбцах. Для каждой извлеченной пары «объект – признак» на пересечении соответствующих ей строки и столбца матрицы ставится 1. Все остальные элементы матрицы, для которых нет соответствий, обнуляются. Ранее на рис. 1 был представлен фрагмент одной из построенных таким образом объектно-признаковых матриц.



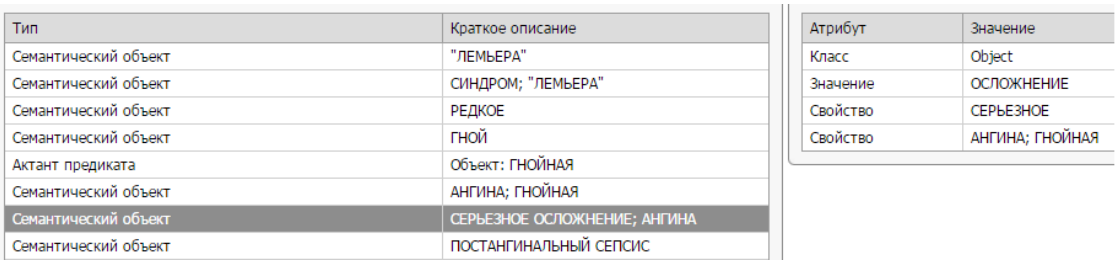
Список объектов (можно выбирать текущий)

Тип	Краткое описание
Семантический объект	ГАЙМОРИТ
Семантический объект	ОСЛОЖНЕНИЕ
Семантический объект	ИНФЕКЦИЯ
Актант предиката	Объект: ИНФЕКЦИОННОЕ
Семантический объект	ЗАБОЛЕВАНИЕ; ИНФЕКЦИОННОЕ
Семантический объект	СЛЕДСТВИЕ; ОСЛОЖНЕНИЕ и ЗАБОЛЕВАНИЕ
Семантический объект	НАПРИМЕР
Семантический объект	СКАРЛАТИНА
Семантический объект	ГРИПП
Семантический объект	ПРОСТУДА
Семантический объект	СИМПТОМ; ГАЙМОРИТ
Семантический объект	ЗАТРУДНЕНИЕ

Текущий объект

Атрибут	Значение
Класс	Object
Значение	СИМПТОМ
Свойство	ГАЙМОРИТ

Рис. 3. Работа программного пакета PullEnti по извлечению объектов и их признаков из текста с описанием симптомов гайморита



Тип	Краткое описание
Семантический объект	"ЛЕМЬЕРА"
Семантический объект	СИНДРОМ; "ЛЕМЬЕРА"
Семантический объект	РЕДКОЕ
Семантический объект	ГНОЙ
Актант предиката	Объект: ГНОЙНАЯ
Семантический объект	АНГИНА; ГНОЙНАЯ
Семантический объект	СЕРЬЕЗНОЕ ОСЛОЖНЕНИЕ; АНГИНА
Семантический объект	ПОСТАНГИНАЛЬНЫЙ СЕПСИС

Атрибут	Значение
Класс	Object
Значение	ОСЛОЖНЕНИЕ
Свойство	СЕРЬЕЗНОЕ
Свойство	АНГИНА; ГНОЙНАЯ

Рис. 4. Работа программного пакета PullEnti по извлечению объектов и их признаков из текста с описанием осложнения гнойной ангины

Разбиение (кластеризация) входной объектно-признаковой матрицы

На рис. 5 приведен фрагмент формального контекста, сформированного на основе текстов медицинской направленности при помощи программного пакета Pullenti и предложенного нами шаблона. Анализируя фрагмент, можно заметить, что признаки, стоящие в парах с объектами, обозначающими болезни, не используются в паре с объектами, обозначающими органы, и наоборот. Например, можно сказать «осложнение гайморита» или «осложнение гриппа», но нельзя сказать «осложнение мозга», и наоборот, можно сказать «отек пазухи» или «отек мозга», но нельзя сказать «отек гриппа». Благодаря наличию таких отличительных признаков мы можем выделить во входной объектно-признаковой матрице группы (кластеры) родственных объектов. Если теперь в каждом кластере убрать нулевые признаки и оставить только те, которые присутствуют хотя бы у одного члена кластера, то мы получим на основе этих кластеров новые менее разреженные объектно-признаковые матрицы меньших размеров.

На рис. 6 показаны результаты иерархической кластеризации объектно-признаковой матрицы, построенной на основе рассмотренного выше формального контекста. Первый кластер, как и ожидалось, содержит названия болезней, а второй – названия органов. Внутри второго кластера в отдельный подкластер выделены объекты «гайморова пазуха» и «печень», которые в формальном контексте встречались вместе с признаком «лечение». Лечение – это пример гибридного признака, который употребляется как с болезнями, так и с органами (сравните: «лечение гепатита» и «лечение печени»).

В данной работе в качестве метода кластеризации мы используем агломеративную кластеризацию по Уорду. При использовании метода Уорда на каждом шаге алгоритма происходит слияние кластеров, которое приводит к минимальному увеличению дисперсии внутри объединенного кластера:

$$\Delta(A,B)=\sum_{i\in A\cup B}\left\|\bar{x}_i-\bar{m}_{A\cup B}\right\|^2-\sum_{i\in A}\left\|\bar{x}_i-\bar{m}_A\right\|^2-\sum_{i\in B}\left\|\bar{x}_i-\bar{m}_B\right\|^2,$$

где A,B – это кластеры до слияния, $\bar{x}_i$ и m – объекты кластеров и их центроиды, $\Delta(A,B)$ – целевая функция, которая называется ценой слияния.

	профилактика	лечение	воспаление	отек	увеличение	осложнение
аппендицит	1	1	0	0	0	1
аппендикс	0	0	1	0	1	0
гайморит	1	1	0	0	0	1
гайморова пазуха	0	1	1	1	1	0
печень	0	1	1	1	1	0
гепатит	1	1	0	0	0	1
мозг	0	0	1	1	1	0
грипп	1	1	0	0	0	1

Рис. 5. Фрагмент формального контекста

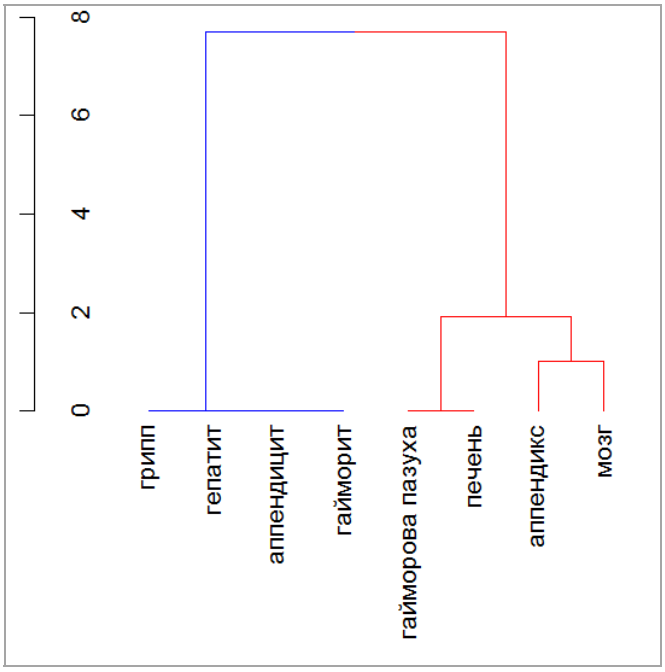


Рис. 6. Кластеры объектов,
выделенные на основе иерархической кластеризации

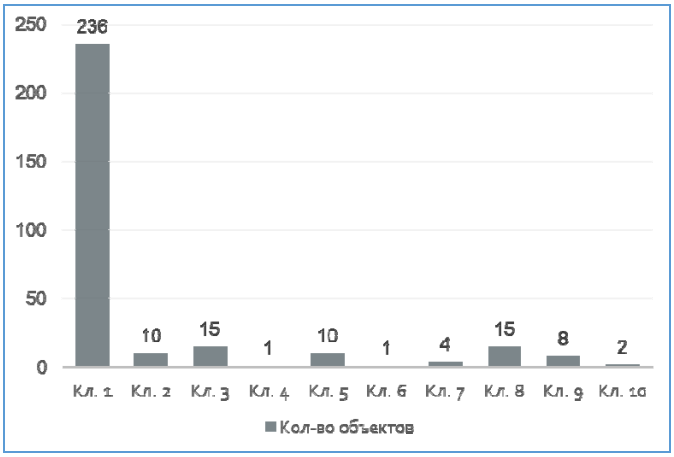


Рис. 7. Пример распределения объектов по кластерам

На рис. 7 представлена типичная диаграмма распределения объектов по кластерам. Типичной ее делает наличие одного гигантского кластера на фоне множества мелких. Эксперименты показывают, что мелкие кластеры, как правило, достаточно устойчивы и действительно состоят из однородных объектов, в то время как гигантский кластер, как правило, представляет собой собрание всех «неясных» объектов, не отнесенных ни к одному из устойчивых кластеров.

Восстановление пропусков
в объектно-признаковых матрицах

Как указывал один из основоположников статистической обработки естественного языка Дж. Ципф, разреженность – это извечная проблема при работе с текстовыми корпусами [8]. Разреженность негативно влияет на результаты кластеризации, искажает их, так что определенная часть терминов в результате такой кластеризации оказывается не в своей категории. Для борьбы с разреженностью данных в этом проекте мы используем подходы на основе слабого машинного обучения. Термин «слабое обучение» возник сравнительно недавно для обозначения методов, в которых обучение, называемое бутстраппингом (англ. bootstrapping – применительно к информационным технологиям «самозагрузка»), ведется на небольшом множестве позитивных примеров [9; 10].

Например, системе, извлекающей из текстов названия болезней, может потребоваться предоставить небольшое количество примеров названий. После этого система ищет в текстах вхождения этих названий, и пытается определить некоторые общие контексты их появления (шаблоны). Затем система по выявленным шаблонам пытается отыскать новые названия болезней, например, в сети Интернет. Процесс обучения является итеративным, что позволяет обнаруживать всё новые шаблоны и новые экземпляры названий. В итоге многократные повторения этого процесса позволят собрать большое количество названий болезней и большое количество шаблонов их употребления в текстах.

В нашем случае в качестве обучающих примеров мы используем самые «сильные» (близко расположенные к центроиду) объекты кластера и определяем контексты их появления в текстах коллекции. В обнаруженном контексте мы заменяем «сильный» объект кластера на «слабый», например на объект, категория которого вызывает у нас сомнение, и пытаемся отыскать вхождения «слабого» объекта в рамках заданного контекста в сети Интернет. Для этой цели в нашей системе разработан поисковый робот, который осуществляет просмотр тематических веб-страниц и ищет вхождения слабого объекта. Мы формируем задания этому роботу по сбору веб-страниц из сети Интернет, описывая схемы, позволяющие произвести правильный парсинг страниц. Описание схем производится в полуавтоматическом режиме: основная структура веб-страницы загружается в систему, и необходимо настроить только некоторые параметры разметки страниц, например указать контейнер основного текста, последовательность страниц для обхода и т. д.

Таким образом, вслед за авторами работы [11] мы используем Интернет как сверхбольшой и сверхценный корпус текстов и оцениваем вероятность использования слабого объекта в контексте, опираясь на количество откликов (hits) поисковой машины на запрос:

$$Score\left(Weak\ Object, Context\right) = \frac{hits\left(Context\ with\ Weak\ Object\right)}{hits\left(Weak\ Object\right)}.$$

В табл. 2 представлены кластеры, которые получились после уменьшения разреженности в исходном наборе данных, состоящем из 119 строк (объектов) и 616 столбцов (признаков). Жирным шрифтом выделены слова, которые, с точки зрения эксперта, являются релевантными основному содержанию кластера, подчеркнуты слова, ошибочно включенные в данный кластер. Не все кластеры, с точки зрения эксперта, объединяют однородные объекты, тем не менее эти результаты демонстрируют принципиальную возможность использования предлагаемого подхода. Ошибочное появление слов в том или ином кластере можно объяснить, во-первых, разреженностью полученной матрицы, во-вторых, неполнотой исходных корпу-

сов текстов, в-третьих, тем, что решение задачи кластеризации неоднозначно и существенно зависит от выбора метода кластеризации и критерия оценки качества.

Таблица 2

Кластеры, полученные после уменьшения разреженности

Номер кластера и тематика	Термины
1 – симптомы, признаки, проявления	"активность" "**действие**" "**образование**" "период" "**повреждение**" "**признак**" "**проявление**" "**развитие**" "**симптом**" "**состояние**" "фаза" "**функция**"
2 – действия общего характера, методы	**"борьба" "введение" "возникновение" "выделение" "выявление" "использование" "контакт" "передача" "повышение" "прием" "применение" "проведение" "работа" "распространение" "снижение" "течение" "увеличение"**
3 – препараты и воздействующие факторы	**"антиген" "вирус" "возбудитель" "кровь" "препарат" "фактор"**
4 – методы медицинского контроля и лечения	"диагностик" **"диагностика" "исследование" "контроль" "лечение" "определение" "профилактика"** "фон"
5 – группирующие слова	**"вид" "группа" "ряд" "система" "тип" "форма" "характер"**
6 – структурные элементы организма	"боль" **"клетка" "орган" "ткань"** "человек" **"эритроцит" "лейкоцит" "бактерия"**
7 – характеристики, проявления симптомов и болезней	**"выраженность" "гиперемия"** "день" "дифтерия" **"длительность"** "зависимость" "заражение" **"исчезновение" "картина" "конец" "момент" "наличие" "нарастание" "отсутствие" "патогенез" "подозрение" "попадание" "появление" "причина" "продолжительность" "проникновение" "случай"** "смерть" **"срок" "стадия" "тенденция" "тяжесть" "уменьшение"** "употребление"
8 – количественные, измеряемые характеристики	"возможность" **"время" "интенсивность" "количество" "концентрация"** "локализация" "место" "особенность" **"показатель"** "результат" **"содержание"** "способность" **"степень" "уровень" "частота" "число"** "этиология"
9 – слова, обозначающие заболевание	**"болезнь" "заболевание" "изменение" "нарушение" "поражение" "процесс"**
10 – области	"кожа" "лицо" **"область"** "организм" **"основа" "очаг" "поверхность" "структура" "часть"**

На рис. 8, 9 представлены результаты иерархической кластеризации объектов, примененной соответственно к кластерам 2 и 8. Как видно из рисунка, объекты внутри кластера можно, в свою очередь, также разделить на семантические подгруппы в соответствии с дифференцирующими признаками.

Извлечение формальных понятий и их иерархий

На рис. 10 представлена решетка понятий, сформированная на основе кластера 6. Как видно из этой решетки, лейкоциты, эритроциты и бактерии – это клетки. Причем, отличие эритроцитов и лейкоцитов от бактерий заключается в том, что они могут оседать, у них есть ядро, и они подвержены анизоцитозу, а отличие бактерий в том, что они могут быть патогенными. Клетка, орган и ткань согласно диаграмме Хассе – это понятия одного уровня. Хотя на самом деле и органы, и ткани состоят из клеток, поэтому клетка должна стоять уровнем ниже, предлагаемый подход позволяет выделять только отношения иерархии “is-a” и не позволяет выделять отношения вида “part-of”.

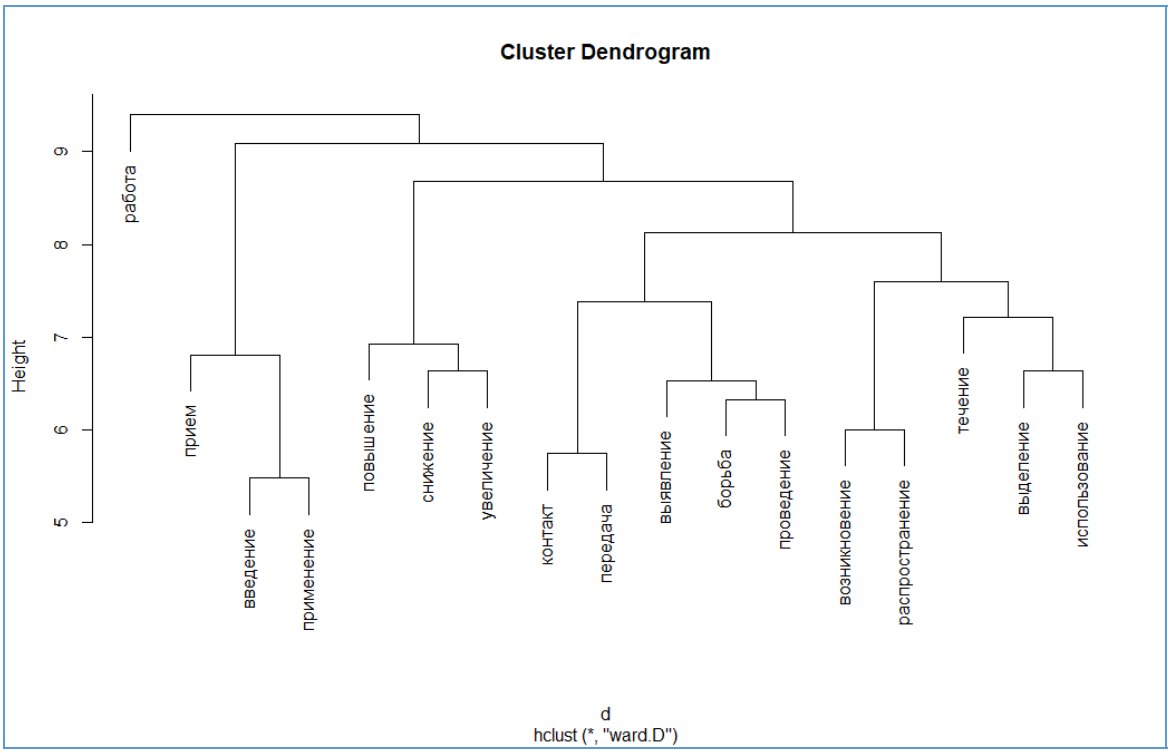


Рис. 8. Результаты повторной кластеризации,
примененной к кластеру объектов, обозначающих действие

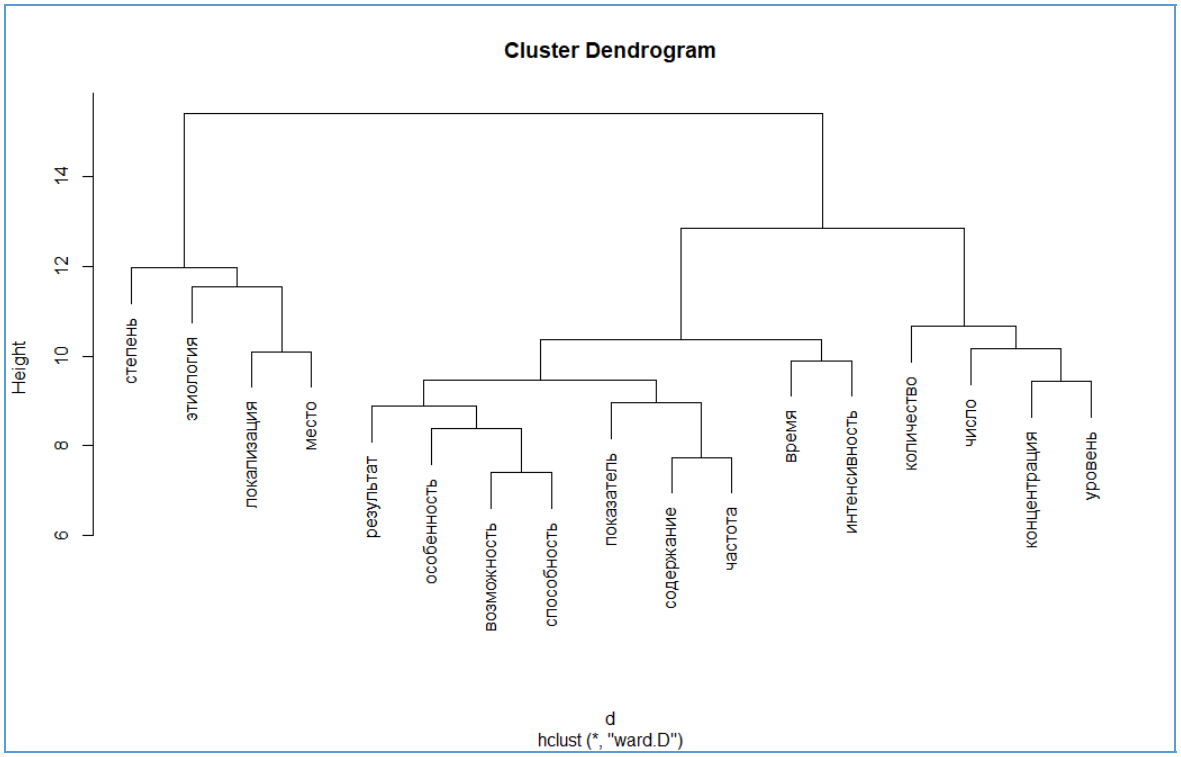


Рис. 9. Результаты повторной кластеризации, примененной к кластеру объектов,
обозначающих количественные, измеряемые характеристики

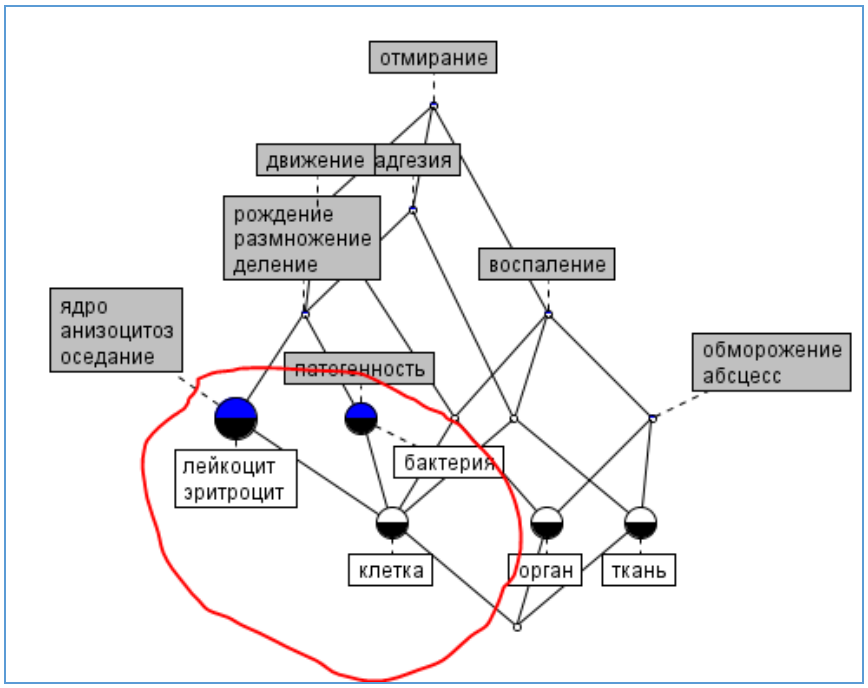


Рис. 10. Решетки понятий для кластера 6

Заключение

Предложенный в статье новый способ извлечения понятий из текстов предметной области, основанный на комбинации анализа формальных понятий и бутстрап-технологии информационного поиска с использованием данных сети Интернет продемонстрировал высокую эффективность при работе с корпусами текстов, в которых специальная терминология сильно разрежена. Таким образом, при правильном выборе исходных шаблонов поиска бутстраппинг, основанный на использовании открытых ресурсов Интернета как ценных источников знаний, превращается в эффективный инструмент поддержки концептуального моделирования.

В своих будущих исследованиях мы планируем апробировать данный подход на текстах других тематик, а также попробовать формировать формальный контекст на основе новых шаблонов, например шаблонов «существительное + существительное», но с использованием предлогов. По-прежнему остается актуальным вопрос определения количества кластеров: пока мы решаем эту задачу эмпирическим путем.

Список литературы

1. *Игнатов Д. И.* Анализ формальных понятий: от теории к практике // Доклады всероссийской научной конференции АИСТ'12 «Анализ изображений, сетей и текстов». 16–18 марта 2012 г. Национальный открытый университет «ИНТУИТ». Екатеринбург, 2012. С. 3–15.

2. *Ganter B., Wille R.* Formal concept analysis: mathematical foundations. Springer Science & Business Media, 2012. 284 p.

3. *Кузнецов О. С., Объедков С. А.* Алгоритмы построения множества всех понятий формального контекста и его диаграммы Хассе // Изв. РАН. Теория и системы управления. 2001. № 1. С. 120–129.

4. *Hwang Y. S., Finch A., Sasaki Y.* Improving statistical machine translation using shallow linguistic knowledge // Computer Speech & Language. 2007. Vol. 21. No. 2. P. 350–372.

5. *Crysmann B. et al.* An integrated architecture for shallow and deep processing // Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002. P. 441–448.

6. PullEnti / К. И. Кузнецов. 2013. URL: http://www.pullenti.ru/Default.aspx (дата обращения 07.01.2018).

7. *Kozerenko E., Kuznetsov K., Morozova Yu., Romanov D.* Semantic Proximity Establishment in the Tasks of Knowledge Extraction and Named Entities Recognition // Proc. of the 2017 Int. Conf. on Artificial Intelligence. 2017. P. 339–344.

8. *Zipf G.* Selective Studies and the Principle of Relative Frequency in Language. Cambridge, 1932.

9. *Nadeau D., Turney P., Matwin S.* Unsupervised Named-Entity Recognition: Generating Gazetteers and Resolving Ambiguity // Advances in Artificial Intelligence. 2006. P. 266–277.

10. *Schapire R. E.* The boosting approach to machine learning: An overview // Nonlinear estimation and classification. New York: Springer, 2003. P. 149–171.

11. *Vieira K. et al.* Finding seeds to bootstrap focused crawlers // World Wide Web. 2016. Vol. 19. No. 3. P. 449–474.

Материал поступил в редколлегию 24.08.2018

A. B. Nugumanova, E. M. Bayburin, M. E. Mansurova, V. B. Barakhnin

[1] *Sarsen Amanzholov East-Kazakhstan State University*
34 Tridtsatoy Gvardeiskoy Divizii Str., Ust-Kamenogorsk, 070002, Republic of Kazakhstan

[2] *Al-Farabi Kazakh National University*
71 al-Farabi ave., Almaty, 050040, Republic of Kazakhstan

[3] *Institute of Computational Technologies SB RAS*
6 Academician Lavrentiev Ave., Novosibirsk, 630090, Russian Federation

[4] *Novosibirsk State University*
1 Pirogov St., Novosibirsk, 630090, Russian Federation

{anugumanova, ebaiburin}@vkgu.kz, mansurova.madina@gmail.com, bar@ict.nsc.ru

AUTOMATIC EXTRACTION OF FORMAL LATTICES FROM MEDICAL TEXTS BASED ON THE COMBINATION OF THE FORMAL CONCEPT ANALYSIS AND BOOTSTRAPPING TECHNOLOGIES

The article considers a new way of concept extraction from the subject domain texts based on combination of formal concept analysis and bootstrap technology of information retrieval. Formal concept analysis is a powerful way of automatically deriving the domain concepts, but it is designed for high quality input data, without missing and inaccuracies. Obtaining such datasets directly from texts is difficult because of the strong sparsity of the text corpora. Accordingly, it seems promising to improve the quality of input data with bootstrapping, a technology that provides an intelligent search for fragmented information on the Internet. In this paper, we show the steps of implementing the way of automatically concept extraction from medical texts based on the filling of blanks in highly sparse matrices of the joint occurrence of terms. The input data for formal concept analysis is represented in the form of an object-feature table that reflects the distribution of attributes over the objects of the domain. The purpose of this paper is to show that with proper selection of initial search patterns, bootstrapping based on the use of open Internet resources as valuable sources of knowledge, turns into an effective tool for supporting conceptual modeling.

Keywords: formal concept analysis, bootstrapping, information extraction, surface linguistic analysis, information retrieval.

References

1. Ignatov D. I. Analysis of formal concepts: from theory to practice. *Reports of the Russian scientific conference AINT'12 "Analysis of images, networks and texts."* March 16–18, 2012. National Open University "INTUIT". Ekaterinburg, 2012, p. 3–15.

2. Ganter B., Wille R. Formal concept analysis: mathematical foundations. Springer Science & Business Media, 2012, 284 p.

3. Kuznetsov O. S., Obedkov S. A. Algorithms for constructing the set of all concepts of a formal context and its Hasse diagram. *Izvestiya RAN. Theory and control systems*, 2001, no. 1, p. 120–129. (in Russ.)

4. Hwang Y. S., Finch A., Sasaki Y. Improving the statistical translation of shallow linguistic knowledge. *Computer Speech & Language*, 2007, vol. 21, no. 2, p. 350–372.

5. Crysmann B. et al. An integrated architecture for shallow and deep processing. *Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics*, 2002, p. 441–448.

6. PullEnti / K. I. Kuznetsov. 2013. URL: http://www.pullenti.ru/Default.aspx (date of access 07.01.2018).

7. Kozerenko E., Kuznetsov K., Morozova Yu., Romanov D. Semantic Proximity Establishment in the Tasks of Knowledge Extraction and Named Entities Recognition. *Proc. of the 2017 Int. Conf. on Artificial Intelligence*, 2017, p. 339–344.

8. Zipf G. Selective Studies and the Principle of Relative Frequency in Language. Cambridge, 1932.

9. Nadeau D., Turney P., Matwin S. Unsupervised Named-Entity Recognition: Generating Gazetteers and Resolving Ambiguity. *Advances in Artificial Intelligence*, 2006, p. 266–277.

10. Schapire R. E. The boosting approach to machine learning: An overview. *Nonlinear estimation and classification*. New York, Springer, 2003, p. 149–171.

11. *Vieira K. et al.* Finding seeds to bootstrap focused crawlers. *World Wide Web*, 2016, vol. 19, no. 3, p. 449–474.

Received 24.08.2018

For citation:

Nugumanova A. B., Bayburin E. M., Mansurova M. E., Barakhnin V. B. Automatic Extraction of Formal Lattices from Medical Texts Based on The Combination of the Formal Concept Analysis and Bootstrapping Technologies. *Vestnik NSU. Series*: *Information Technologies*, 2018, vol. 16, no. 4, p. 140–152. (in Russ.) DOI 10.25205/1818-7900-2018-16-4-140-152