УДК 004.4’413
DOI: 10.15587/1729-4061.2018.147978

Разработка информационной технологии выделения терминов из документов на естественном языке

А. Б. Кунгурцев, С. Л. Зиноватная, Я. В. Поточняк, М. А. Кутасевич

Показано, що словники предметних областей широко використовуються на різних етапах створення і експлуатації програмних продуктів. Процес створення словника, особливо виділення термінів, досить трудомісткий та вимагає високої кваліфікації експерта. Проведено дослідження по виявленню найбільш важливих характеристик багатослівних термінів, таких як: ймовірності присутності в документі термінів, що містять різну кількість слів; розташування іменників в багатослівних термінах; можливу кількість іменників в багатослівних термінах. Проаналізовано контекст використання термінів та визначено можливі межі термінів в тексті. Запропоновано процедуру попереднього групування документів, що дозволяє уникнути «втрати» термінів, що входять в короткі документи. Визначено залежність помилок при виділенні термінів від розміру аналізованого документа.

Запропоновано математичну модель представлення терміна, що заснована на визначенні безлічі ланцюжків слів, згрупованих близько опорного слова – іменника. Фільтрація ланцюжків виробляється в залежності від частоти їх входження в текст на основі зіставлення нормалізованих уявлень багатослівних термінів.

Розроблено механізми заповнення словника предметної області новими записами і коригування існуючих у міру аналізу вхідного документа. Запропоновано рішення щодо коригування частоти появи термінів на основі виявлення міжфразових зв'язків. Всі процеси і моделі об'єднані в єдину інформаційну технологію створення словника предметної області. Проблема визначення тлумачень термінів в даній роботі не розглядається, оскільки вимагає окремого рішення. Розроблено програмний продукт, що дозволяє в значній мірі автоматизувати процес виділення термінів з текстових документів. Результати апробації запропонованих рішень показали відсутність «загублених термінів» і, як результат, скорочення часу виділення термінів з текстів обсягом в 10000 слів на 1.5 години за рахунок звільнення експерта від аналізу вихідного документа. Результати дослідження можуть бути використані на різних етапах створення і експлуатації програмних продуктів

Ключові слова: словник предметної області, багатослівний термін, морфологічний розбір, математична модель терміна, текстовий документ

1. Введение

Словари предметных областей (СПО) широко используются при проектировании программных продуктов [1]. В частности, при определении ролей членов команды разработчиков [2]; при построении словарей данных [3,

4]; в задачах выбора и кластеризации материализованных представлений баз данных [5, 6]. На основе СПО создаются должностные инструкции и многие другие документы. Для построения СПО производится анализ различных текстов, используемых в конкретной предметной области. Это могут быть приказы, отчеты, договора, инструкции и другие документы, в достаточной степени отражающие деятельность конкретной организационной системы. Основным этапом построения СПО является выделение терминов из текстов. При этом под термином понимают не только отдельные слова, но и устойчивые словосочетания или многословные термины (МТ) в конкретной предметной области. Выделение МТ вручную требует длительного труда специалиста высокой квалификации [7, 8]. Поэтому исследования, направленные на автоматизацию процессов выявления МТ для построения СПО, являются актуальными.

2. Анализ литературных данных и постановка проблемы

В работе [9] предложен программный пакет LEXTER для извлечения терминов. Представляет интерес то, что термины формируются на основе выделения существительных. Связанные с ними слова определяются эмпирическими правилами, что ограничивает область применения пакета французским языком. Рассматриваемый в работе [10] статистический метод выделения терминов применим для славянских языков. Однако авторы решают задачу выделения терминов в контексте кластеризации документов и поиска контрастных терминов. Это приводит к появлению большого числа ложных терминов. В работе [11] поставлена задача извлечения не только отдельных ключевых слов, но и словосочетаний, объединяемых по частотным характеристикам. Однако решение предложено для построения иерархической кластеризации документов, когда нужно определить не все термины и выделенные термины содержат не более двух слов. Представляет интерес исследования по выделению ключевых фраз [12], которые можно использовать при формировании толкования терминов в СПО. Однако с точки зрения выделения терминов ключевая фраза требует дальнейшего анализа. В работе [7] предложен метод автоматизированной предварительной группировки текстов и выделения терминов по частотным характеристикам и упрощенным синтаксическим правилам. Это позволило выделять термины из двух и частично трех слов. Однако формирование термина по схеме «существительное + прилагательные» не позволило выделять все многословные термины, а двойной проход по документам снижал производительность. В работе [13] глубокий синтаксический и семантический анализ документов на естественных языках. Однако предлагаемые модели не доведены до такой степени формализации, которая позволяет использовать их в прикладных задачах. Интересное решение по снижению трудоемкости выделения ключевых слов путем организации параллельных вычислений предложено в работе [14]. Однако предложенный алгоритм применим только для выделения однословных терминов и теряет эффективность при небольшом количестве документов, что характерно для узких предметных областей.

Таким образом, задача выделения терминов для построения СПО имеет ряд нерешенных проблем, а именно:

– исследование характеристик терминов, позволяющих сформулировать требования к выделению их из текста (количество слов, расположение и тип опорных слов, ограничения);

– предварительная группировка текстов, обеспечивающая обнаружение терминов в документах малого объема;

– разработка технологии, обеспечивающей выделение терминов, содержащих произвольное количество слов;

– создание программного продукта, реализующего предложенную технологию и позволяющего апробировать принятые решения.

3. Цель и задачи исследования

Целью исследования является сокращение времени и повышение качества выделения терминов из документов в узкой предметной области.

Для достижения цели были сформулированы следующие задачи:

– определить характеристики терминов, влияющие на технологию выделения их из текста;

– создать информационную технологию выделения терминов из текста, включающую предварительную группировку документов;

– создать программный продукт и оценить качество выделения терминов.

4. Определение характеристик многословных терминов

Для автоматизации процесса выделения МТ потребовалось выявить ряд характеристик МТ, влияющих на технологию этого процесса. Определяемые характеристики используют понятие «опорное слово» – существительное, входящее в МТ. Для работы с МТ потребовались следующие характеристики:

– возможное количество слов, входящих в термин;

– расположение «опорных слов» в МТ;

– возможное количество «опорных слов» в МТ;

– определение слов и знаков препинания, которые ограничивают МТ.

Словарь предметной области используется для проектирования и сопровождения программного продукта. Поэтому для исследования были выбраны текстовые документы из разных областей техники и прикладных наук на русском, украинском и белорусском языках. Для каждой предметной области было выделено по 200 терминов.

На рис. 1 показано усредненное распределение вероятностей вхождения в многословный термин определенного количества слов. Разброс значений, определяемый конкретной предметной областью, не превысил 1–2 %.

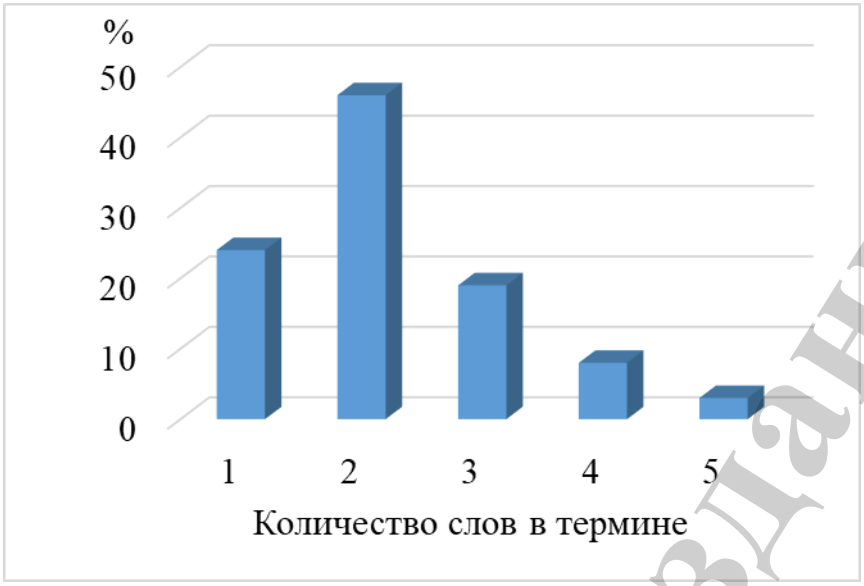

Рис. 1. Вероятности появления термина, содержащего одно и более слов

На рис. 2 приведены результаты анализа расположения опорного слова в многословном термине. В качестве опорного слова выбирались существительные, например, для термина «информационная система» опорным словом является «система». Если в термине оказалось более одного существительного как, например, в термине «реляционные базы данных», то каждое из них было отнесено к соответствующей категории («базы» – посредине, «данных» – справа).

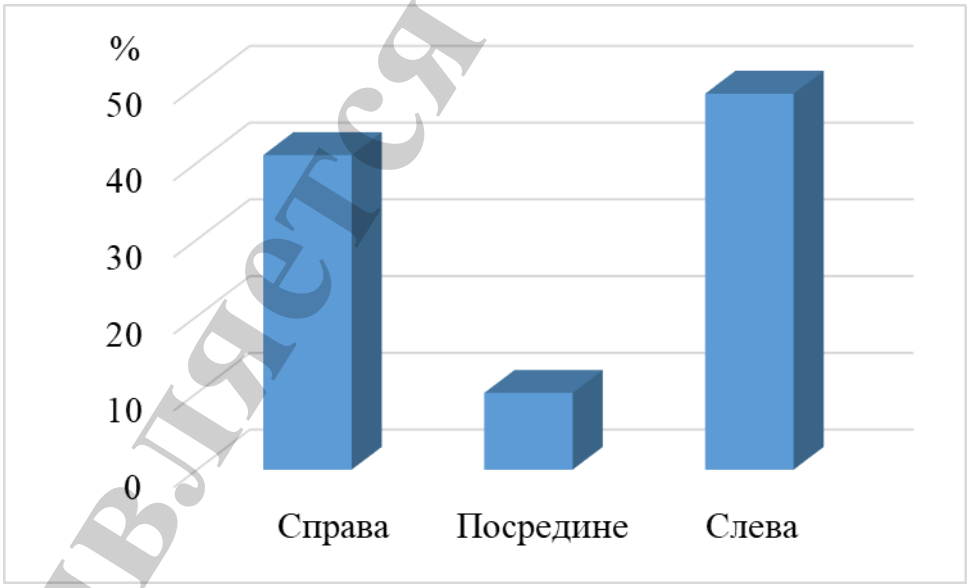

Рис. 2. Вероятности расположения опорного слова в многословном термине

На рис. 3 показана вероятность появления нескольких опорных слов (существительных) в МТ. Из приведенных данных следует, что вероятность появления нескольких существительных (опорных слов) в термине высокая. Поэтому способ выделения термина по существительному и согласованному с

ним прилагательному [7] приводит к большим погрешностям, а более глубокий анализ связи слов требует сложного синтаксического анализа. Это подтверждает необходимость искать более эффективный способ выделения МТ.

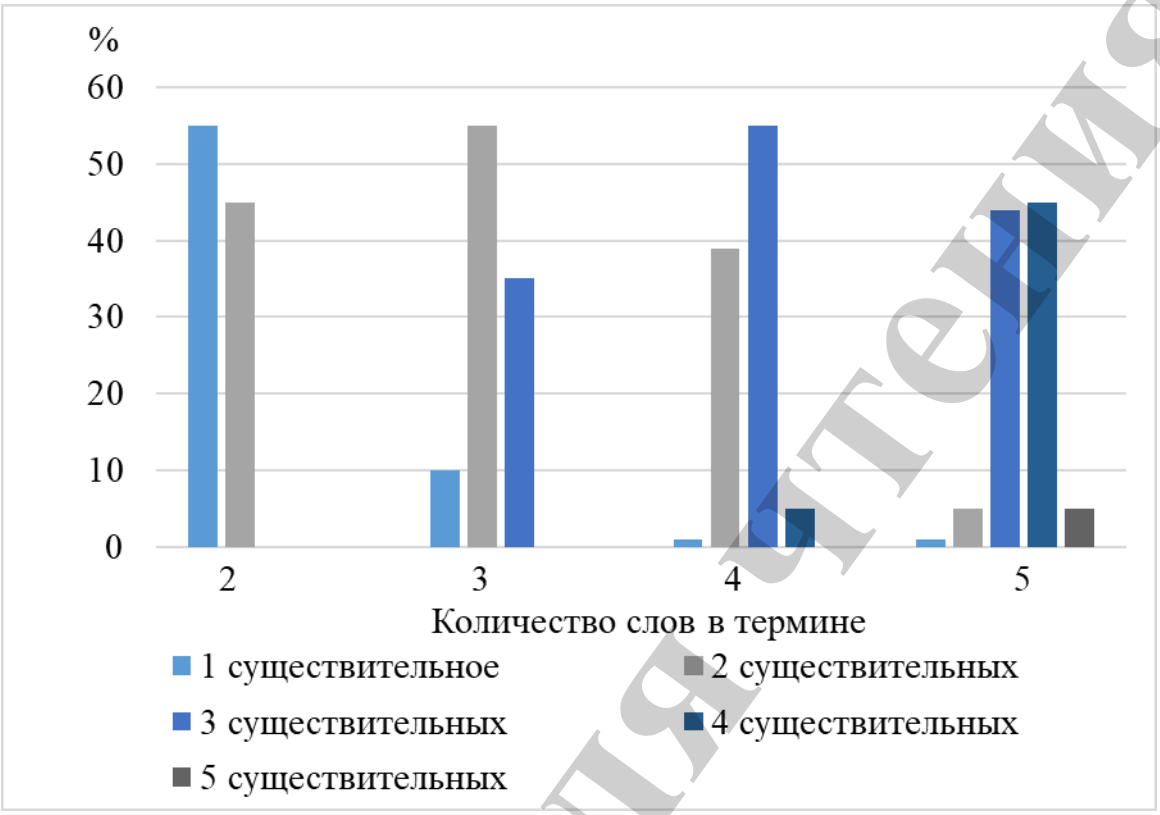

Рис. 3. Вероятность появления существительных в МТ

В табл. 1 приведены результаты определения возможных границ МТ.

Таблица 1
Возможные границы вхождения МТ в текст

№	Ограничение слева	Вхождение в МТ	Ограничение справа
1	Пробел	Входит	Пробел
2	, пробел	Входит	,
3	– пробел	Входит	Пробел –
4	: пробел	Не входит?	:
5	; пробел	Не входит	;
6	. пробел	Не входит	.
7	? пробел	Не входит	?
8	! пробел	Не входит	!
9	) пробел	Не входит	Пробел (
10	» пробел	Не входит	Пробел «
11	Местоимение пробел	Не входит	Пробел местоимение

Случай, когда запятая входит в МТ (№ 2), оказался единственным из 1000 проанализированных МТ («лица, принимающие решения»).

В соответствии с результатами исследования сделаны следующие выводы:

– многословный термин может быть представлен не более чем пятью словами;

– расположение и количество опорных слов в многословном термине может быть любым;

– последовательность слов, входящих в многословный термин, должна быть ограничена слева и справа знаками препинания или местоимениями.

5. Технология выделения терминов из текста

Технология предусматривает ряд этапов:

– подбор и группировка документов, отражающих предметную область;

– преобразование формата документов;

– морфологический разбор анализируемого текста, выделение существительных;

– определение возможных МТ на основе опорных слов;

– выявление межфразовых связей и замена ссылок на термины терминами;

– подсчет количества вхождений МТ в текст;

– корректировка словаря терминов путем поиска вхождений одних терминов в другие.

Предлагаемая технология применима для наиболее распространенных европейских языков. Для славянских языков ввиду наличия склонений по падежам, согласования по родам и сравнительно сложных правил образования множественного числа, для сравнения терминов дополнительно введен механизм использования нормализованной формы их представления.

5. 1. Предварительная группировка документов.

В процессе анализа предметной области (ПРО) для определения требований к разрабатываемому программному продукту системному аналитику приходится иметь дело с множеством документов. Эти документы могут представлять различные аспекты деятельности исследуемой организации. Выделение терминов из всего множества документов как единого целого может привести к недооценке тех терминов, которые сосредоточены в отдельных небольших документах.

Обработка каждого документа по отдельности при малом объеме некоторых из них может не обеспечить накопления статистики. Для определения влияния размера документа на качество выделения терминов было проведено исследование множества документов разного объема. Полагалось, что если некоторое словосочетание встретилось в документе один раз, то при автоматизированном способе выделения терминов оно не будет опознано как термин. Если эксперт посчитает это словосочетание термином, то это указывает на потенциальную ошибку автоматизированного поиска терминов. На основании анализа 100 документов с различным количеством слов была получена зависимость вероятности ошибки определения термина от размера документа (рис. 4).

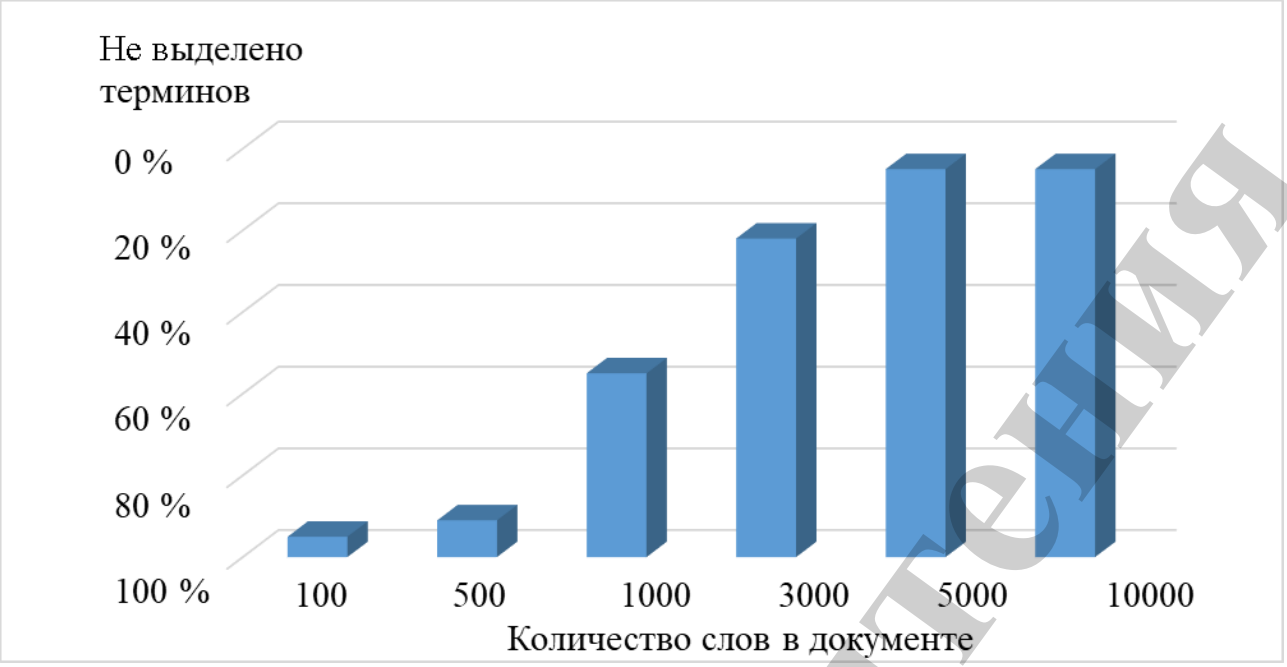

Рис. 4. Зависимость ошибки в выделении термина от размера документа в словах

В работе [7] предложено группировать документы на основе нормализованного расстояния между ними, что предусматривает проведение частичного морфологического разбора. В данной работе введём понятие объема v_i некоторого документа D_i. Если оказалось, что v_i<5000, то в соответствии с рис. 4 будем считать, что документ D_i должен быть объединен с другими документами в некоторый объединенный документ T_x для достижения группой объема не менее чем 5000 слов:

$$T_x = \{D_i\} \mid \sum_{i=1}^{n} v_i \geq 5000.$$

Задачу подбора документов для группы можно поручить эксперту в предметной области или системному аналитику.

5. 2. Преобразование форматов документов

Известные анализаторы текстов [15] принимают на входе документы в формате .txt, поэтому требуется соответствующее преобразование:

$$T_x \Rightarrow T_{txt}.$$

Подобное преобразование выполняет любой текстовый редактор.

5. 3. Математическая модель многословных терминов

Предлагаемая модель рассматривает однословный термин как частный

случай многословного термина. В результате обработки текста T_{txt} должен быть получен список терминов. Назовем этот список словарем, поскольку в дальнейшем при добавлении в этот список толкований терминов, он становится словарем ПРО. На этапе выделения терминов словарь представим в виде множества записей:

$$D = \{r_i\} i = 1, n. \quad (1)$$

Каждая запись имеет вид:

$$r = < tm, lsn, nf, q >, \quad (2)$$

где *tm* – множество вариантов представления термина, *lsn* – список опорных слов (существительных), входящих в термин, в нормализованном виде, *nf* – нормализованное представление термина, *q* – количество вхождений термина в документ.

Введение нормализованной формы представления термина необходимо только для славянских языков. В качестве примера приведены три предложения на русском английском и французском языках.

*Мы работаем с **реляционными базами данных**.*
*Нами внесены изменения в **реляционную базу данных**.*
***Реляционная база данных** содержит множество таблиц.*
*We work with **relational databases**.*
*We made changes to a **relational database**.*
*The **relational database** contains a set of tables.*
*Nous travaillons avec les **bases de données relationnelles.***
*Nous faisons les changements dans la **base de données relationnelle***
*La **base de données relationnelle** contient la multitude de tableaux.*

Все предложения содержат термин «реляционная база данных». Однако варианты его представления на русском языке существенно отличаются один от другого, тогда как на английском и французском языках эти отличия минимальны. Поэтому для сравнения терминов из текстов на неславянских языках можно успешно применить нечеткое сравнение строк (например, используя расстояние Левенштейна), в то время как для текстов на славянских языках предложено использовать нормализованную форму представления термина.

Представление одного термина множеством вариантов *tm* также используется для славянских языков, поскольку позволяет в конце анализа выбрать правильное представление многословного термина, содержащего несколько опорных слов. Каждый элемент множества *c* состоит из одинаковой последовательности слов. Различаются элементы падежами и единственным либо множественным числом соответствующих слов. Табл. 2 иллюстрирует использование множества вариантов представления одного термина в славянских языках.

Таблица 2

Варианты представления многословного термина

Номер варианта	Термин		
	Белорусский язык	Украинский язык	Русский язык
1	рэляцыйнымі базамі дадзеных	реляційними базами даних	Реляционными базами данных
2	рэляцыйную базу дадзеных	реляційну базу даних	реляционную базу данных
3	рэляцыйная база дадзеных	реляційна база даних	Реляционная база данных
	Нормализованная форма представления термина		
	Реляцыйныя база дадзенае	реляційний база дане	реляционный база данное

Нормализованная форма представления *nf* является единой для всех вариантов представления термина. В работе [13] предложено сравнивать содержание текстов при помощи специального лингвистического процессора. Использование нормализованной формы позволяет сравнивать многословные термины при помощи очень простой процедуры четкого сравнения строк, что существенно сокращает время обработки текста.

Для неславянских языков множество *tm* из (2) будет содержать один элемент (термин), а *nf* – этот же термин.

Список опорных слов термина *lst* при завершении формирования словаря *D* позволит выбрать наиболее подходящий вариант представления термина *tm*.

В соответствии с диаграммой на рис. 1, в многословный термин может входить до 5 слов. В соответствии с диаграммой на рис. 2, опорное слово в многословном термине может занимать любую позицию. Поэтому предложено сформировать все возможные группы слов относительно опорного слова. С целью сокращения количества возможных групп в соответствии с табл. 1 определено множество видов левых и правых границ МТ:

$$B = \{":",";",".","?","!","(",")","<<",">>",\backslash pron\}. \quad (3)$$

Предложено формировать возможные термины как последовательности из 5, 4, 3, 2 и одного слова, содержащих, как минимум, одно опорное слово. Предварительно предположим, что в последовательность войдет только одно опорное слово.

Представим фрагмент текста *S* в виде последовательности элементов:

$$e_1,...,e_l,..,e_m. \quad (4)$$

Элементом может быть отдельное слово или знак препинания. Каждое слово представлено последовательностью букв *W* (непосредственно из текста),

множеством *A* нормализованной формой представления *nf* (результат работы анализатора):

$$e = < W, A, nf > . \quad (5)$$

Определим атрибуты, которые будут необходимы для определения границ МТ, а также учета межфразовых связей [16]. Пусть *A1* представляет часть речи, *A2* – число, *A3* – род, *A4* – лицо, *A5* – падеж.

Знаки препинания представляются только своим написанием $e = < W, \varnothing, * > .$

Пусть некоторый элемент является опорным словом $e_0 = <W, A, q>$, где *A1=noun* (существительное), *q* – количество появления термина в тексте *S*.

Сформулируем правила составления последовательностей слов:

– последовательность формируется из элементов, расположенных рядом друг с другом;

– опорное слово обязательно входит в последовательность;

– число элементов в последовательности не должно быть более 5 и менее 1 (знаки препинания, входящие в последовательность, не учитываются);

– последовательность может быть ограничена слева или справа от опорного слова, если некоторым элементом предложения e_i при условии, что $e_j \in B$.

Пусть в некотором тексте имеется последовательность элементов:

$$e_{-5}e_{-4}e_{-3}e_{-2}e_{-1}e_0e_1e_2e_3e_4e_5,$$

где e_0 – опорное слово.

Тогда возможными последовательностями слов (без учета ограничений) будут:

$$\begin{bmatrix} e_{-4}e_{-3}e_{-2}e_{-1}e_0 \\ e_{-3}e_{-2}e_{-1}e_0e_1 \\ e_{-2}e_{-1}e_0e_1e_2 \\ e_{-1}e_0e_1e_2e\ e_3 \\ e_0e_1e_2e_3e_4 \\ e_{-3}e_{-2}e_{-1}e_0 \\ ..., e_{-1}e_0 \\ e_0e_1e_2e_3e_4 \\ ..., e_0e_1 \end{bmatrix}. \quad (6)$$

Предлагается формула для определения количество возможных комбинаций:

$$K = \sum_{i=0}^{i\le 5-2} (5-i) = 14. \qquad (7)$$

Учтем возможные границы для комбинаций. Пусть некоторый элемент $e_j \in B$. Тогда все комбинации, в которые входят элементы с индексами $i \le j$, исключаются из дальнейшего анализа. Формула для определения количества возможных комбинаций при ограничении слева имеет вид:

$$K_l = 14 - \sum_{i=1}^{5-j} (5 - j + i - 1). \qquad (8)$$

Рассмотрим более общий случай, когда в группу может войти более одного опорного слова. Пусть в некотором тексте имеется последовательность элементов:

$$e_i \ldots . e_j^* \ldots e_k^* \ldots e_l,$$

где e_j^* и e_k^* – опорные слова. Тогда, при условии, что:

$$k - j \ge 5. \qquad (9)$$

Формируемые последовательности слов в качестве терминов будут содержать по одному опорному слову. Если $k–j<5$, то, используя ранее изложенную методику образования последовательностей слов отдельно для опорного слова e_j^* и для опорного слова e_k^*, получим повторяющиеся последовательности. Например, для фрагмента предложения:

$$e_{=4} e_{-3} e_{-2} e_{-1} e_0^* e_1 e_2 e_3^* e_4 e_5 e_6 e_7 .$$

На основе e_0^* получим следующие последовательности с двумя опорными словами:

$$\begin{bmatrix} e_{-1} e_0^* e_1 e_2 e_3^* \\ e_0^* e_1 e_2 e_3^* e_4 \\ e_0^* e_1 e_2 e_3^* \end{bmatrix}.$$

И на основе e_3^* получим те же последовательности с двумя опорными словами:

$$\begin{bmatrix} e_{-1}e_0^*e_1e_2e_3^* \\ e_0^*e_1e_2e_3^*e_4 \\ e_0^*e_1e_2e_3^* \end{bmatrix}.$$

Ниже будет показано, как исключить повторяющиеся последовательности слов в словаре.

Количество возможных последовательностей слов при наличии в последовательности нескольких опорных слов зависит от числа опорных слов, но не может превышать *K* из (6) в расчете на одно опорное слово.

5. 4. Включение последовательности слов в словарь

Каждая последовательность слов *E*, полученная после учета ограничений, должна быть представлена записью (2) в словаре (1). Для этого определяем её нормализованную форму E_{nf}. Введем обозначение принадлежности некоторой последовательности слов словарю $E \in_e D$. Если:

$$r_i \mid r_i \in D \wedge r_i.nf = E_{nf}$$

то комбинация слов уже присутствует в словаре. В этом случае увеличиваем количество вхождений на 1 ($r_i.q$: $r_i.q$+1) и проверяем вхождение *E* в $r_i.tm$. Если $E \in r_i.tm$, то добавляем новый вариант термина в множество вариантов $r_i.tm = r_i.tm \cup \{E\}$.

Если $E \notin_e D$, то образуем новую запись в словаре:

$$r =< E, lsn, E_{nf}, 1>,$$

где *lsn* будет содержать все существительные (опорные слова, выделенные на этапе построения последовательностей) из *E* в нормализованном виде.

5. 5. Учет межфразовых связей

Основным критерием отбора терминов является частота их появления в анализируемом тексте. Межфразовые связи возникают, если некоторый термин в последующих предложениях заменен местоимением, порядковым числительным и т. п., например, в предложении «*Жесткий диск является основным запоминающим устройством для большинства персональных компьютеров.»* словосочетание *«Жесткий диск»* можно определить, как термин. В следующем предложении «*Обычно он характеризуется емкостью и количеством оборотов в минуту.»* термин *«Жесткий диск»* заменен местоимением «он». Если связь между предложениями не будет обнаружена, то будет определено только одно появление термина *«Жесткий диск»*. В настоящем исследовании были использованы результаты, полученные в работе [16], где приведены алгоритмы выявления межфразовых связей. Пусть

некоторый элемент предложения e_i оказался анафорой (заменой или ссылкой) ранее обнаруженного термина $e_i \rightarrow r_i.t$, тогда следует увеличить число вхождений $r_i.t$ в текст:

$$r_i.q := r_i.q + 1.$$

Здесь знак «:=» обозначает присваивание.

5. 6. Корректировка словаря

Для каждого термина необходимо ввести нижнюю границу *Be* количества вхождений термина в *G*:

$$\forall r \in G \mid r_i.m \geq Be.$$

Минимальное значение нижней границы – *Be*=2. При этом значении *Be* некоторая последовательность слов, выделенная в соответствии с (8), повторно встретилась в тексте. Для больших текстов значение *Be* можно увеличить. Рекомендуется эту операцию поручить эксперту в предметной области. Последовательно увеличивая значение *Be*, нужно зафиксировать момент, когда в словаре ещё сохраняются все важные для данной предметной области термины.

В результате анализа документа могут появиться термины, которые входят в другие термины. Вопрос о сохранении таких терминов в словаре, или исключении их из словаря, зависит от их самостоятельного использования в тексте. Процедура корректировки словаря предусматривает сравнение записей. Если:

$$r_i.nf \in_t r_j.nf \wedge r_i.q = r_j.q,$$

то запись r_i исключается из словаря.

Если:

$$r_i.nf \in_t r_j.nf \wedge r_i.q > r_j.q,$$

то следует проанализировать $\Delta = r_i.q - r_j.q$. Если $\Delta \geq Be$, то запись r_i не исключается из словаря.

После определения терминов, которые должны войти в словарь. Необходимо выбрать для каждого термина один из вариантов его представления в множестве *tm*. Для этого введем понятие «главное слово» в термине. Существует ряд признаков, которые отличают его от других слов:

– оно должно быть существительным (обязательно);

– обычно оно занимает первое место среди других существительных входящих в термин;

– обычно именно его варианты написания (изменения по падежам и числу) задают различные варианты представления термина в *tm*.

Таким образом, процесс выделения варианта представления термина предусматривает следующую последовательность действий.

Определяем количество элементов множества *tm*.

Если $|tm|=1$, то имеется только один вариант представления термина, который будет представлен в словаре.

Если $|tm|=k \wedge k>1$, то определяется число опорных слов в списке *lsn*.

Если $|lsn|=1$, то имеется только одно опорное слово $w_1 \in lsn$ в термине. Его позиция j в вариантах представления термина определяется на основании позиции этого слова в *r..nf*:

$$w_1 = w_j \mid w_j \in r..nf. \tag{10}$$

Далее из каждого варианта представления термина $tm_i \in tm$ выделяется слово $w_{i,j}$ в позиции j и сравнивается с нормализованным представлением.

Если:

$$w_{i,j} = w_1, \tag{11}$$

то из множества *tm* удаляются все элементы кроме tm_i, то есть $tm=\{tm_i\}$.

Если условие (11) не выполнено, то:

$$tm = \{tm_1\}, \tag{12}$$

и проблему формулировки определения термина решает эксперт.

Если $|lsn|=l \wedge l>1$, то в определении термина имеется несколько опорных слов. Для каждого опорного слова $w_p \in lsn | p=1, l,$ определяется его позиция j в вариантах представления термина в соответствии с (10).

Далее из каждого варианта представления термина $tm_i \in tm$ выделяется слово $w_{i,j}$ в позиции j и сравнивается с нормализованным представлением.

Если $w_{i,j}=w_p$, то из множества *tm* удаляются все элементы кроме tm_i, то есть $tm_i \in tm$ и цикл поиска наилучшего варианта представления термина завершается. В противном случае $p=p+1$ и цикл продолжается. Если наилучшее представление не было найдено, то принимается решение в соответствии с (12) и проблему формулировки определения термина решает эксперт.

6. Создание программного продукта и оценка качества выделения терминов

Для реализации предложенной технологии и моделей был создан программный продукт TermsSelect. Схема обработки документа представлена на рис. 5.

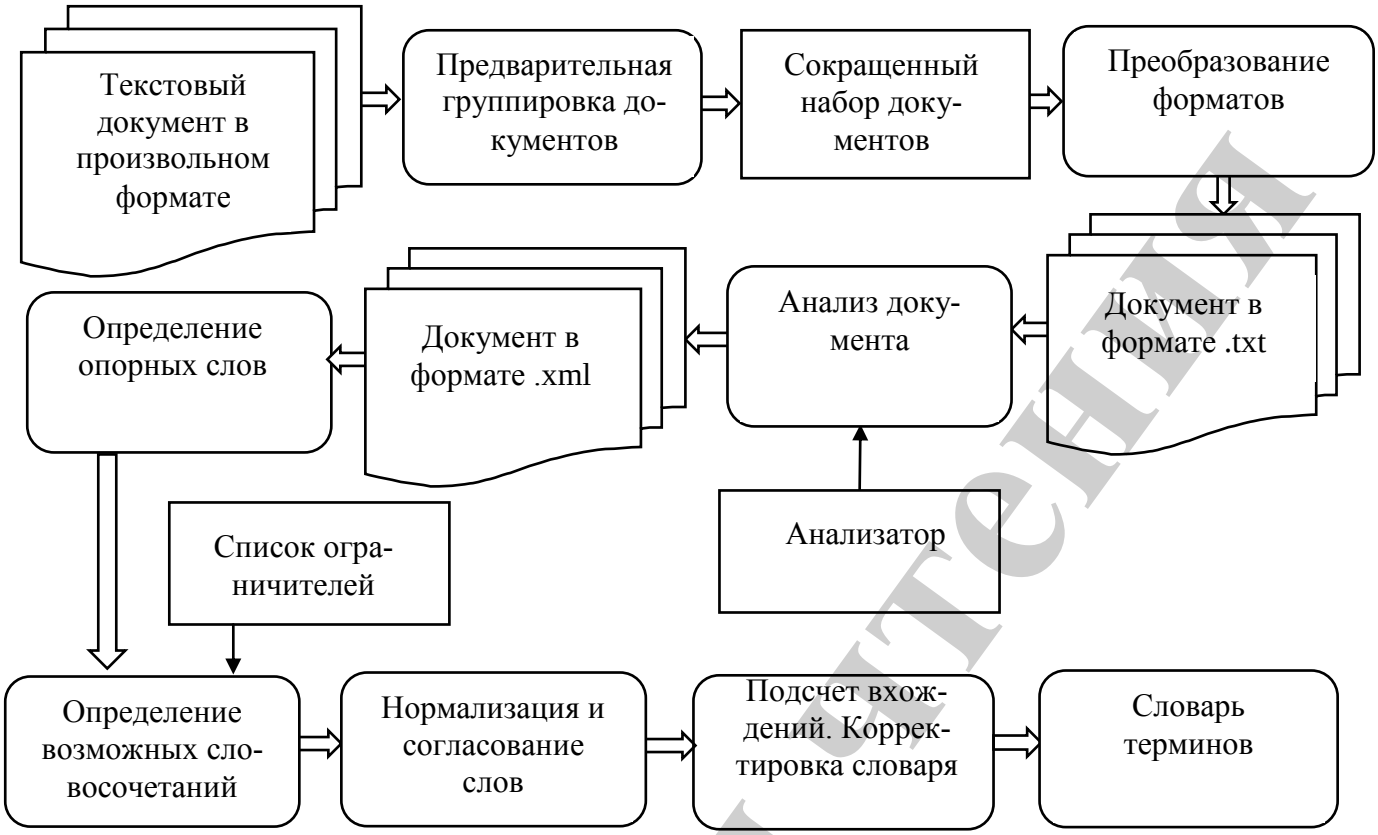

Рис. 5. Функциональная схема программы TermsSelect

На рис. 6 представлено окно, позволяющее эксперту отредактировать список найденных в тексте терминов. Термины были получены в результате анализа 15 текстов по теме «Материалы и технологии изготовления керамических изделий» общим объемом около 20000 слов [17]. Расстановку терминов определяет первое существительное. Редактированию подлежит содержимое первой колонки «Термин». Кроме этого, эксперт может удалить строку из таблицы, либо вписать в таблицу новый термин.

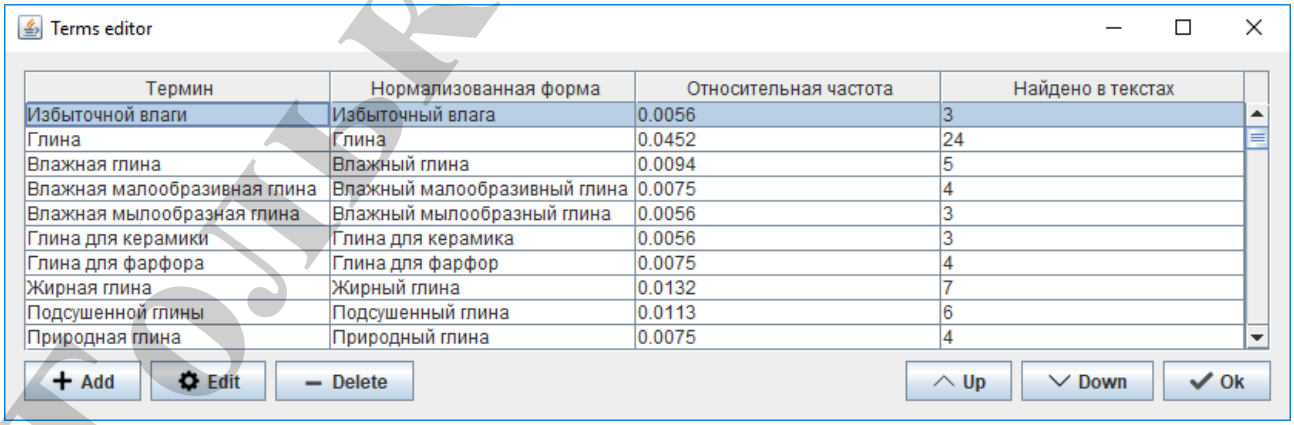

Рис. 6. Редактирование словаря терминов

Целью испытаний программного продукта была сравнительная оценка новой и ранее существовавшей технологии по временным характеристикам и

качеству выделения терминов. Под качеством понимался процент ошибок первого рода («лишние» термины) и второго рода («утерянные» термины) от общего количества обнаруженных терминов.

Для проведения испытаний предложенной технологии и программного продукта были использованы тексты из различных областей науки и техники. В результате проведенных экспериментов установлено, что при использовании TermsSelect среднее время выделения терминов из документа объемом 10000 слов составило 15,6 секунд. Хронометраж работы эксперта по выделению терминов и их частотных характеристик «вручную» дал результат около 10 часов. Упрощенную задачу – только выделение терминов эксперт выполнил в течение 1.5 часа. При выделении терминов программой были обнаружены «лишние термины». Они составили около 5% от выделенных терминов. При этом «утерянных терминов» не было выявлено. Следует отметить, что удаление «лишних терминов» не требует специальной процедуры, поскольку во всех случаях список выделенных терминов просматривается экспертом.

Для сравнения было проведено испытание программного продукта DictionaryCreator, предложенного в [7]. Здесь время выделения терминов составило 12.4 секунды для документа объемом 10000 слов. Однако количество «утерянных терминов» составило 22 % (в основном термины из трех и более слов). Определение «утерянных терминов» весьма трудоемкая процедура, которую можно выполнить только вручную. Таким образом, при незначительном увеличении времени обработки текстов удалось получить существенное повышение качества выделения многословных терминов.

7. Обсуждение результатов исследования с точки зрения скорости и качества выделения терминов

Существенное сокращение количества «утерянных терминов» при высокой скорости обработки исходных текстов объясняется двумя основными решениями:

– предложенным способом формирования потенциальных терминов как всех допустимых цепочек слов, расположенных около опорных слов;

– предварительной группировкой коротких документов.

Представление термина множеством цепочек слов позволяет определить термины как подмножество цепочек, которые повторяются в тексте. Такой принцип может быть использован для большинства естественных языков и требует только морфологический анализ. Группировка коротких документов на период анализа позволяет найти термины, встречающиеся в одном документе один раз. Предложенное решение требует от эксперта только редактирования термина, включенного в словарь. Существующие решения по определению частот одиночных слов отличаются высоким быстродействием, однако оставляет эксперту очень много работы, связанной с анализом исходных текстов. Способы выделения термина как существительного с относящимися к нему прилагательными не охватывает всего многообразия терминов. Быстродействие такого способа соизмеримо с предложенным в данной работе,

однако большое число «утерянных терминов» также требуют от эксперта работы с исходным текстом.

Проведенные исследования были ограничены славянскими и наиболее распространенными европейскими языками, для которых можно ввести понятие опорного слова. Они не могут бать распространены, например, на вьетнамский и другие языки, подобные китайскому.

К недостаткам исследования следует отнести представление выделенного термина в виде, который в ряде случаев требует редактирования экспертом. Попытки представить многословный термин в окончательном виде без применения известных трудоемких способов генерирования текста пока не увенчались успехом.

Кроме общепринятого понятия термина в текстах, представляющих узкую предметную область, могут использоваться специфические аббревиатуры и названия (программ, процессов, машин и т. д.), которые также можно отнести к терминам. Выделение такого рода терминов требует формализации понятия «аббревиатура», «название» и является продолжением данного исследования. Словарь предметной области должен содержать толкование терминов, которое в настоящее время выполняется вручную. Автоматизация этого процесса предполагает поиск соответствующих источников информации и выделение подходящих фрагментов текста. Решение этой задачи также требует дальнейших исследований.

8. Выводы

1. Определены такие параметры терминов, как возможное количество входящих в него слов, возможное количество и позиции существительных в термине, а также возможные ограничители цепочки слов, входящих в термин. Результаты исследования необходимы для построения математической модели термина.

2. Разработана информационная технология выделения терминов из текстовых документов, содержащая группировку документов; математическую модель термина, позволяющую выделить его из предложения; корректировку частот появления терминов на основе выявления межфразовых связей и вхождения одних терминов в другие. Технология позволяет выделять термины без детального синтаксического анализа предложения, что существенно сокращает время обработки документа.

3. Создан программный продукт TermsSelect, реализующий предложенную технологию. На вход подавались текстовые документы в любых общепринятых форматах. Для выделения частей речи и получения нормализованной формы представления слов использовались подключаемые анализаторы текста, находящиеся в свободном доступе. Максимальная длина цепочки слов была установлена равной пяти. Задачей эксперта являлось только редактирование терминов. В качестве аналога был взят ранее разработанный программный продукт DictionaryCreator, выделяющий термины как существительные и синтаксически связанные с ними прилагательные. Сравнительные испытания продуктов на одинаковых текстах показали, что при почти одинаковых затратах

времени на обработку текста TermsSelect обнаружил все термины, а DictionaryCreator 78 % терминов. Поиск «утерянных терминов» был оценен в 1.5 часа работы эксперта. Таким образом, достигнутое повышение качества выделения терминов позволило существенно сократить общее время выделения терминов.

Література

1. Избачков Ю. С., Петров В. Н. Информационные системы: учеб. Питер, 2011. 544 с.
2. Liubchenko V., Sulimova I. Examining the attributes of transitions between team roles in the software development projects // Eastern-European Journal of Enterprise Technologies. 2017. Vol. 1, Issue 3 (85). P. 12–17. doi: https://doi.org/10.15587/1729-4061.2017.91597
3. Best Practices for Data Dictionary Definitions and Usage Version 1.1. 2006. URL: https://s3.us-west-2.amazonaws.com/org-pnamp-assets/prod/best_practices_for_data_dictionary_definitions_and_usage_version_1.1_2006-11-14.pdf
4. 10 Ways Data Dictionary Increases Software Developers Productivity. URL: https://dataedo.com/blog/ways-data-dictionary-increases-software-developers-productivity
5. Novokhatska K., Kungurtsev O. Application of Clustering Algorithm CLOPE to the Query Grouping Problem in the Field of Materialized View Maintenance // Journal of Computing and Information Technology. 2016. Vol. 24, Issue 1. P. 79–89. doi: https://doi.org/10.20532/cit.2016.1002694
6. Novokhatska K., Kungurtsev O. Developing methodology of selection of materialized views in relational databases // Eastern-European Journal of Enterprise Technologies. 2016. Vol. 3, Issue 2 (81). P. 9–14. doi: https://doi.org/10.15587/1729-4061.2016.68737
7. Кунгурцев А. Б., Поточняк Я. В., Силяев Д. А. Метод автоматизированного построения толкового словаря предметной области // Технологический аудит и резервы производства. 2015. Т. 2, № 2 (22). С. 58–63. doi: https://doi.org/10.15587/2312-8372.2015.40895
8. Califf M., Mooney R. J. Bottom-up relational learning of pattern matching rules for information extraction // Journal of Machine Learning Research. 2003. Vol. 4. P. 177–210.
9. Bourigault D. Surface grammatical analysis for the extraction of terminological noun phrases // COLING '92 Proceedings of the 14th conference on Computational linguistics. 1992. P. 977–981. DOI: https://doi.org/10.3115/993079.993111
10. Метод контрастного извлечения редких терминов из текстов на естественном языке / Бессмертный И. А., Нугуманова А. Б., Мансуровас М. Е., Байбурин Е. М. // Научно-технический вестник информационных технологий, механики и оптики. 2017. Т. 17, № 1. С. 81–91. doi: https://doi.org/10.17586/2226-1494-2017-17-1-81-91
11. Попова С. В., Ходырев И. А. Извлечение ключевых словосочетаний // Научно-технический вестник Санкт-Петербургского государственного универ-

ситета информационных технологий, механики и оптики. 2012. № 1 (77). С. 67–71.

12. Hasan K. S., Ng V. Automatic keyphrase extraction: a survey of the state of the art // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014. P. 1262–1273. doi: https://doi.org/10.3115/v1/p14-1119

13. Вавіленкова А. І. Критерії аналізу логіко-лінгвістичних моделей речень природної мови // Вісник Національного Технічного Університету "Харківський політехнічний інститут". Серія: Нові рішення в сучасних технологіях. 2017. № 7 (1229). С. 118–122. doi: https://doi.org/10.20998/2413-4295.2017.07.16

14. Реализация алгоритма извлечения ключевых слов из текстов предметной области на основе модели MapReduce / Бессмертный И. А., Каримов А. Т., Новоселов А. О., Нугуманов А. Б. // Труды VIII Международной научно-практической конференции "Современные информационные технологии и ИТ-образование". 2013. С. 617–624.

15. Программный пакет синтаксического разбора и машинного перевода. URL: https://www.cognitive.ru/

16. Учет межфразовых связей при автоматизированном построении толкового словаря предметной области / Кунгурцев А. Б., Гаврилова А. И., Леонгард А. С., Поточняк Я. В. // Информатика и математические методы в моделировании. 2016. № 2. С. 173–183.

17. Материалы и технология изготовления керамических изделий. URL: http://art-con.ru/node/233