DOI 10.31891/2307-5732-2018-267-6(2)-202-208
УДК 004.4:81'374

А. Б. КУНГУРЦЕВ, Я. В. ПОТОЧНЯК, А. Г. ЛИПИНСКАЯ, Л. С. ЖИРО
Одесский национальный политехнический университет

ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ ВЫДЕЛЕНИЯ АББРЕВИАТУР ДЛЯ СЛОВАРЕЙ ПРЕДМЕТНЫХ ОБЛАСТЕЙ

Показано, что словари предметных областей широко используются на различных этапах создания и эксплуатации программных продуктов. Модель выделения термина для аббревиатур определяет цепочку слов в зависимости от выявленного типа аббревиатуры. Каждой аббревиатуре соответствует многословный термин. Выделение этого термина – нетривиальная задача, поскольку единого способа введения в текст документа аббревиатуры не существует. Поэтому предложено решение исследовать документ с точки зрения выделения различных типов аббревиатур и определения вероятности их появления. Разработан программный продукт, позволяющий в значительной степени автоматизировать процесс выделения аббревиатур из текста.

Ключевые слова: словарь предметной области, многословный термин, анафора, аббревиатура.

O. B. KUNGURTSEV, IA. V. POTOCHNIAK, A. G. LIPINSKA, L. S. ZHIRO
Odessa National Polytechnic University

INFORMATION TECHNOLOGY EXTRACTION ABBREVIATURES FOR DICTIONARY OF SUBJECT AREAS

Dictionaries of subject areas are widely used at various stages of creating and operating software products. The aim of the study is to develop a method for identifying abbreviations and proper names in the texts of documents from a specific subject area. The term allocation model for abbreviations defines a chain of words depending on the type of abbreviation identified. Each abbreviation corresponds to a verbose term. Selection of this term is a non-trivial task, since there is no single way to introduce an abbreviation into the text of a document. Therefore, it was proposed to investigate the document in terms of identifying different types of abbreviations and determining the likelihood of their occurrence. A software product has been developed that allows automating the process of extracting abbreviations from the text to a significant degree. The results of approbation of the proposed solutions showed a significant reduction in time for the process of identifying abbreviations and proper names from natural language texts with a minimum number of errors.

Keywords: dictionary of subject area, multi-word term, anaphora, abbreviation.

Введение

Словари предметных областей (СПО) широко используются при создании программных продуктов, проектировании баз данных и знаний, в процессе эксплуатации различных организационных систем [1]. Основной принцип построения СПО для некоторой предметной области – определения частотных характеристик вхождения терминов в тексты документов, принадлежащих этой предметной области [2, 3]. Известны работы по уточнению частоты появления терминов путем выявления анафор [4], когда термин представлен местоимением, числительным или другими частями речи. Наряду с анафорами термины также могут быть представлены аббревиатурами. Статья СПО не может содержать только аббревиатуру. Необходимо поместить в словарь и соответствующий ей термин. Найти термин по аббревиатуре можно с помощью словарей сокращений русского, украинского, английского и других языков. Однако такие словари не привязаны к узкой предметной области, поэтому могут предлагать десятки терминов для одной аббревиатуры. Так, например, для аббревиатуры «ООП» в словаре [5] предложено около 200 толкований. Из сказанного следует, что автоматизация процесса выделения аббревиатур и соответствующих им многословным терминов – актуальная задача построения словарей узкой предметной области.

Анализ литературных данных и цель работы

В работах [5, 6] рассматриваются вопросы классификации и различные аспекты использования аббревиатур, однако не предложено решений по выделению их из текстов.

В работе [7] предложено решение для нахождения полного названия журнала по его аббревиатуре. Узкая специализация полученного решения не позволяет его использовать для определения полных форм аббревиатур в разных предметных областях.

В работах [8, 9] на основании определения частот соседей для слов определяется мера связности слов, что позволяет предложить вероятные полные формы сокращений. Достоинством такого метода является его универсальность, недостатком – высокая трудоемкость.

В работах [10, 11] предложено исходный текст представлять в виде множества тем, которые образуются множеством входящих в них с разной вероятностью слов. Найденная схожесть частей текста может быть использована как представление сокращения. Указанный подход обычно предлагает множество решений, возможно, с близкими вероятностями, что предусматривает много работы для эксперта на стадии формулировки термина.

В работе [12] сделана попытка объединения нескольких рассмотренных ранее подходов для нахождения полной формы сокращений.

Во всех рассмотренных работах (кроме [4]) предложены универсальные решения для определения сокращений без выделения аббревиатур, как частного случая общей задачи. Результатом является использование весьма трудоемких алгоритмов и во многих случаях нечеткий результат анализа. Поскольку при составлении СПО интерес представляют только аббревиатуры и только в редких случаях специальные сокращения, для которых необходимо получить детерминированные определения, то возникает проблема

разработки специализированного метода определения полной формы для аббревиатур. Таким образом целю исследования является сокращение времени на нахождение полной формы для аббревиатур в виде единственного многословного термина.

Для достижения поставленной цели предложено решить следующие задачи.

1. Определить орфографические правила образования аббревиатур и выделения специальных сокращений.

2. Определить вероятностные характеристики появления в текстах аббревиатур различных типов.

3. Разработать математическую модель аббревиатуры.

4. Создать программный модуль и выполнить апробацию принятых решений.

1. Классификация аббревиатур, на основе их орфографии

Аббревиатуры принято подразделять на инициальные и сложносокращенные [6]. В первом случае аббревиатура составляется из первых букв последовательности слов (для нас эта последовательность является многословным термином). Во втором случае в аббревиатуру могут быть включены не только первые, но и другие буквы сокращаемых слов. Классические сложносокращенные аббревиатуры типа «зарплата» в конкретной предметной области рассматриваются как однословные термины и в дальнейшем нами анализироваться не будут.

С точки зрения определения термина, соответствующего аббревиатуре, предлагается следующая классификация аббревиатур:

A. Инициальная аббревиатура на языке документа заключенная в круглые скобки; слова термина разделены только пробелами. Например, «…словарь предметной области (СПО)…».

B. Инициальная аббревиатура на языке документа заключенная в круглые скобки; некоторые слова термина объединены знаком тире либо между словами встречается запятая, либо некоторое слово термина заключено в круглые скобки. Например, «…объектное – ориентированное программирование (ООП)…» или «…и помогает лицам, принимающим решения (ЛПР), находить…» или «Современные исследования и разработки по созданию перспективных интеллектуальных (экспертных) систем поддержки принятия решений (ИСППР) …».

C. Инициальная аббревиатура с элементами сложносокращенных слов на языке документа заключенная в круглые скобки. Например, «…система автоматизированного проектирования (САПР)…». Здесь «ПР» представляет собой сокращение слова «проектирование».

D. Инициальная аббревиатура на иностранном языке заключенная в круглые скобки; слова термина разделены только пробелами. Например, «unified process (UP)».

E. Инициальная аббревиатура на иностранном языке заключенная в круглые скобки; некоторые слова термина объединены знаком тире либо между словами встречается запятая, либо некоторое слово термина заключено в круглые скобки. Например, «high – pressure cylinder (HPC)».

F. Инициальная аббревиатура на иностранном языке заключенная в круглые скобки совместно с многословным термином на иностранном языке. Например, «Вывод на основе прецедентов (CBR – Case – Based Reasoning) является…»

G. Инициальная аббревиатура на иностранном языке не заключенная скобки; термин на иностранном языке располагается перед аббревиатурой. Например, «…как On – Line Analysis Processing, или OLAP…».

H. Инициальная аббревиатура, заключенная в круглые скобки, на иностранном языке; термин – на языке документа. Например, «…гликемический индекс (GI)…»

I. Неинициальная аббревиатура, представляющая химическое соединение. Например, «…гелиальгинат кальция (Alg2Ca)…».Отличительной особенностью такой аббревиатуры является использование прописных букв, которые могут занимать любую позицию, кроме первой.

2. Вероятностные характеристики появления в текстах аббревиатур различных типов

Предложенная классификация была использована для анализа 100 документов из области техники, экономики, пищевых производств, энергетики, механики, экологии, материаловедения, прикладной физики. Для каждого вида аббревиатуры подсчитывалось количество её появления в документах.

Результаты представлены на диаграмме (рис. 1). Из диаграммы видно, что первые пять типов аббревиатур составляют почти 90% от общего количества аббревиатур в исследованных текстах. Процесс выделения терминов для этих типов аббревиатур будет подробно рассмотрен ниже.

Пусть Sn представляет некоторое предложение:

$$Sn = e_1 e_2 ... e_i ... e_n ,$$

где e_i – элемент предложения (слово, либо знак препинания).

Определим слово W как последовательность символов

$$W = s_1 s_2 ... s_j ... s_k , \quad (1)$$

где символ может быть некоторой буквой l из множества букв $l \in mL$ или цифрой d из множества цифр $d \in mD$.

Определим операцию выделения символа из слова:

$$s_j = W[j],$$

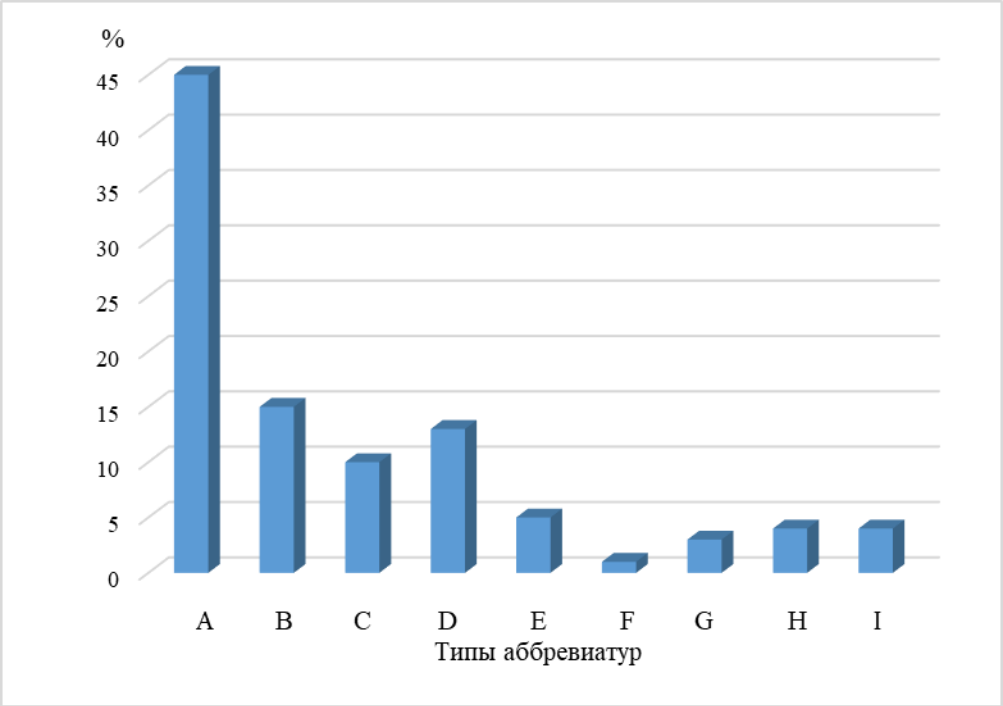

Рис. 1. Вероятности появления аббревиатур различных типов

3. Математическая модель выделения аббревиатур

И отношение принадлежности некоторого символа слову:

$$s_j \in W,$$

Каждая буква характеризуется изображением *li*, размером *is* (строчная, прописная) и алфавитом *al* (кириллица, латиница):

$$l =< li, ls, al >,$$

где *ls* – может принимать два значения low (строчная буква) и cap (прописная буква);

al – может принимать значения la (латиница) и ki (кириллица).

Будем считать, что слово должно начинаться с буквы.

Определим знаки препинания, используемые внутри предложения множеством:

$$Pm = \{",", ":", ";", "(", ")", "–"\}.$$

Выделение инициальных буквенных аббревиатур (типы А и В)

Полагая, что аббревиатура может содержать только буквы, запишем условие первого появления аббревиатуры в тексте.

Пусть $W_m = s_1 s_2 \ldots s_j \ldots s_k$ – некоторое слово, где m – номер слова, как элемента предложения. Если:

$$e_{m-1} = "(" \wedge e_{m+1} = ")" \wedge \forall s_j \mid s_j.ls = cap. \quad (2)$$

То можно считать, что W_m представляет собой аббревиатуру. Обозначим её Ab.

Следующей задачей является определение текста Ta, соответствующего аббревиатуре. Полагаем, что Ta расположен слева от открывающей скобки аббревиатуры и между его элементами отсутствуют какие – либо знаки препинания (тип А) аббревиатуры). Тогда:

$$Ta = e_p \ldots e_{p+(k-1)}, \quad (3)$$

где $p = m - (k+1)$.

Операция по определению Ta будет успешной, если:

$$\forall e_i (i = p, p + (k-1)) \mid e_i[1].li = Ab[j].li (j = 1, k), \quad (4)$$

В тесте, определяющем аббревиатуру, в соответствии с типом В классификации может находиться знак запятой.

Если $e_i \in Ta \wedge e_i = ","$, то элемент e_i не должен учитываться при расчете p для определения Ta и в соответствии с (3) в Ta включается слово e_{p-1}:

$$Ta = e_{p-1} \ldots\ldots e_{p+(k-1)}.$$

Также в тексте термина может использоваться знак тире, связывающий два слова. Если некоторый элемент предложения e_i содержит знак тире – $e_i[j]="–"$, то его следует считать двумя словами и соответственно уменьшить значение p на 1 ($p = p-1$).

В тексте термина, определяющего аббревиатуру, могут присутствовать слова, заключенные в круглые скобки. Эти слова поясняют другие слова термина и не представлены буквами в аббревиатуре. Поэтому слова в скобках не должны учитываться при расчете p для определения Ta. В отличие от аббревиатуры первая буква первого слова в скобках является строчной). Сформулируем условие выделения фрагмента текста, заключенного в круглые скобки. Пусть:

$$e_1 e_2 ... e_i ... e_j ... e_n .$$

Представляет собой некоторый фрагмент текста, расположенный слева от аббревиатуры. Тогда если:

$$e_i = "(" \wedge e_j = ")" \wedge e_{i+1}[1].is \neq cap .$$

То все слова, заключенные между e_i и e_j, не должны учитываться при расчете длины Ta. Тогда:

$$p = m - (k+1) + j - i + 1 .$$

Выделение инициальных буквенных аббревиатур с элементами сложносокращенных слов

Если условие (4) не соблюдается, то можно предположить, что в аббревиатуре присутствует сложносокращенное слово, то есть она относится к типу С.

Поскольку первой букве аббревиатуры всегда должно соответствовать первое слово термина, которое начинается с этой же буквы, выделяем эту букву – $l = W_m[1]$. Затем определяем количество повторений первой буквы в аббревиатуре – n. В соответствии с п.1 при расчете длины термина учитываем слова, заключенные в круглые скобки, знаки препинания и тире. Начиная от слова с номером $m-1$ ищем n слов, таких что:

$$l = e_j[1].li(j = m-1, m-k) .$$

Если n таких слов не обнаружено, то можно предположить, что одна из букв l в аббревиатуре W_m входит в сокращение не на первой позиции. В этом случае уменьшаем число $n := n-1$ и продолжаем поиск слов. Если в результате оказалось что $n = 0$, то поиск расшифровки аббревиатуры прекращается. В противном случае переходим ко второму этапу.

На втором этапе необходимо убедиться, что имеется возможность определить Ta в соответствии с W_m. Для этого в Ta необходимо найти слова, соответствующие буквам $s_2 ... s_j ... s_k$ из аббревиатуры. Запускаем цикл поиска слов ($i = 1$):

Если $e_{n-i}[1] = s_{i+1}$, то $i := i+1$ пока соблюдается условие $i \leq: k$.

Если $e_{n-i}[1] \neq s_{i+1}$, то нужно искать равенство:

$$e_{n-i}[j] = s_{i+1} . \quad (5)$$

Для $j = 2, j \leq d$, где d – количество букв в слове e_{n-i}. Если такое равенство не будет обнаружено, то поиск расшифровки аббревиатуры прекращается. В противном случае аббревиатура W_m и соответствующий ей текст Ta помещаются в словарь предметной области в качестве термина.

Выделение инициальных буквенных аббревиатур на иностранных языках

Приведенные ниже алгоритмы могут использоваться, если язык документа – кириллица (русский, украинский, белорусский, болгарский, сербский языки).

Термины для аббревиатур типов D и E выделяются аналогично выделению терминов для аббревиатур типов 1) и 2).

Рассмотрим способ выделения термина для аббревиатуры типа F. Аббревиатура, составлена из латинских букв, имеет расшифровку на языке документа и иностранном языке. В круглые скобки заключена аббревиатура и её расшифровка на иностранном языке. Расшифровка на языке документа приведена перед скобками. Например, «Вывод на основе прецедентов (CBR – Case – Based Reasoning) является…».

Для выявления такого варианта аббревиатуры предлагается следующий алгоритм:

1) Определяется слово W_m, которое может быть аббревиатурой. Для этого проверяются следующие условия:

$$e_{m-1} = "(" \wedge \forall s_j \in W_m \mid s_j.ls = cap \wedge s_j.al = la .$$

Для $j = 1, j \leq k$, где k – количество входящих в аббревиатуру букв.

2) По правилам расшифровки инициальных буквенных аббревиатур анализируется текст,

расположенный после аббревиатуры до закрывающей круглой скобки.

3) Производится автоматизированный перевод текста, заключенного в круглые скобки.

4) Выделяется фрагмент текста перед круглыми скобками, содержащий k слов.

5) Если фрагмент текста и перевод с допустимой вероятностью совпадают, то в словарь заносится аббревиатура, её расшифровка на иностранном языке и на языке документа.

В соответствии с типом G аббревиатуры сокращение может не заключаться в круглые скобки. Например, «…как On – Line Analysis Processing, или OLAP…».

Для выявления такого варианта аббревиатуры предлагается следующий алгоритм:

1) Определяется слово W_m, которое может быть аббревиатурой. Для этого проверяются следующие условия:

$$e_{m-1} = \forall s_j \in W_m \mid s_j.ls = cap \wedge s_j.al = la .$$

Для $j = 1, j \le k$, k – количество входящих в аббревиатуру букв.

2) В пределах предложения, в котором обнаружена предполагаемая аббревиатура W_m, слева от W_m определяется последовательность слов, составленных из латинских букв. Предполагается, что эта последовательность будет представлять толкование аббревиатуры Ta.

3) По правилам, изложенным в пункте «Выделение инициальных буквенных аббревиатур», производится анализ соответствия Ta аббревиатуре.

Для аббревиатур типа H и I возникают большие затруднения с точным определением термина. Они связаны с тем, что для них нет соответствия начальных букв в словах термина и в аббревиатуре. А для типа I также нет соответствия между количеством знаков в аббревиатуре и количеством слов в термине. Поэтому для типов H и I принято решения в качестве заготовки для термина предоставить фрагмент текста слева от аббревиатуры. Предполагается, что этот текст в дальнейшем будет редактироваться экспертом.

Пусть $e_1 e_2 ... e_i ... e_n$, – некоторое предложение и $e_i = s_1 s_2 ... s_j ... s_k$ – элемент предложения. Если, как минимум два символа из e_i соответствуют условию:

$$\left.\begin{array}{l} s_l \in e_i \wedge s_l \in Lm \wedge s_l.ls = cap \wedge s_l.al = la \\ s_m \in e_i \wedge s_m \in Lm \wedge s_m.ls = cap \wedge s_m.al = la \end{array}\right\}. \quad (6)$$

То можно считать, что e_i является аббревиатурой. Тогда в качестве термина Ta можно принять:

$$Ta = e_1 e_2 ... e_{i-1}.$$

Если $k \le 2 \times (i-1)$, в остальных случаях:

$$Ta = e_{i-2\times k} ... e_{i-1}$$

В соответствии с условием (6) может быть выделено и собственное имя, например, программного продукта, однако это не приведет к ошибке, поскольку собственные имена также могут быть терминами.

4. Апробация принятых решений

В работе [13] разработан программный продукт, позволяющий автоматизировать процесс выделения терминов из текстовых документов. Используя программный продукт из работы [13], была реализована модель выделения термина для аббревиатур, которая определяет цепочку слов в зависимости от выявленного типа аббревиатуры. Программный продукт TermsSelect был расширен модулем – «выделение аббревиатур». Схема обработки документа представлена на рис. 2.

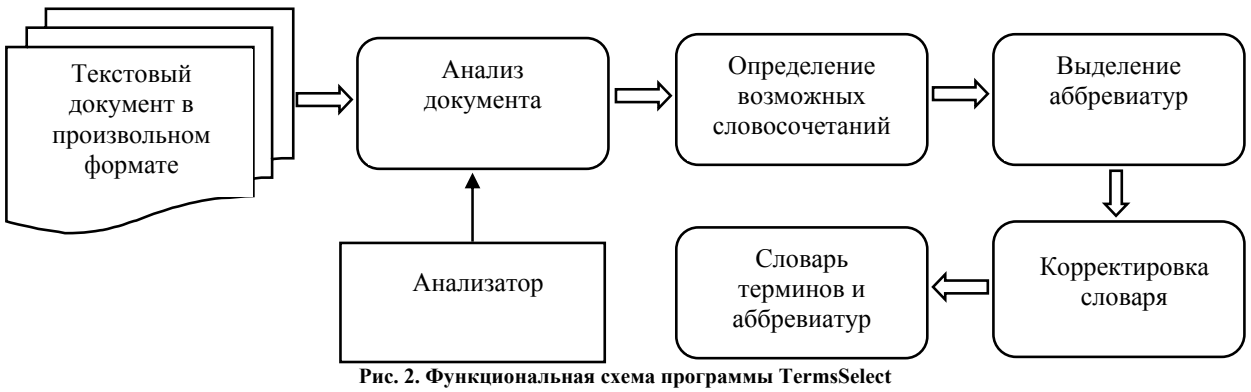

Рис. 2. Функциональная схема программы TermsSelect

На рис. 3 представлено окно, позволяющее эксперту отредактировать термин, определяющий аббревиатуру типа C.

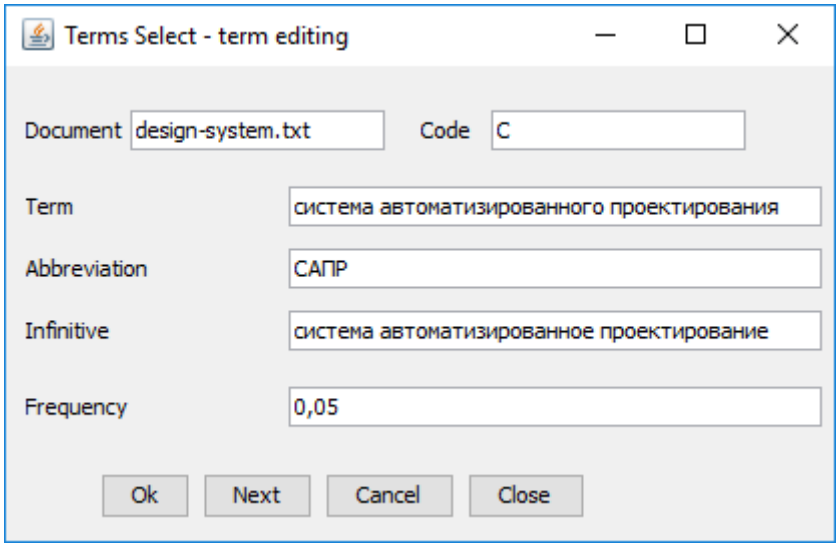

Рис. 3. Просмотр и редактирование термина

Для проведения испытаний предложенной модели были использованы тексты из различных областей науки и техники. Испытание программного продукта, предложенного в работе [13] и добавленного модуля для выделения аббревиатур, показало, что при использовании TermsSelect среднее время выделения термина и определяющую его аббревиатуру, из документа объемом 10000 слов, составило 16,3 секунд. В тех же условиях время выделения терминов экспертом составило 2 часа. При этом ошибок не было выявлено. Таким образом, при обработке текстов удалось получить существенное повышение качества выделения многословных терминов и их аббревиатур.

Заключение

1. Предложена классификация аббревиатур с точки зрения их орфографии, что позволило создать детерминированные процедуры выделения соответствующих терминов.
2. Проведенные статистические исследования показали, что предложенная классификация позволяет выявлять около
3. Разработана математическая модель выделения аббревиатур, позволившая формализовать как процесс выявления аббревиатуры, так и процесс определения её толкования.
4. Создан программный продукт, позволяющий в значительной степени автоматизировать процесс выделения аббревиатур из текста.

Литература

1. Избачков Ю. С. Информационные системы : учебник для вузов / Ю. С Избачков, В. Н. Петров. – Питер, 2011. – 544 с.

2. Кунгурцев А. Б. Метод автоматизированного построения толкового словаря предметной области [Электронный ресурс] / А. Б. Кунгурцев, Я. В. Поточняк, Д. А. Силяев // Технологический аудит и резервы производства – 2015. – № 2/2(22). – С. 58 – 63. – Режим доступа : http://nbuv.gov.ua/UJRN/Tatrv_2015_2%282%29__12.

3. Кунгурцев О. Б. Побудова словника предметної області на основі автоматизованого аналізу текстів українською мовою / О. Б. Кунгурцев, С. В. Ковальчук, Я. В. Поточняк, М. В. Широкоступ // Технічні науки та технології – 2016. – № 3 (5). – С. 164–174.

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

5. Суперанская А. В. Общая терминология: Вопросы теории. Аббревиация в терминологии / А. В. Суперанская, Н. В. Подольская, Н. В. Васильева. – Изд. 6-е. – М. : Книжный дом «ЛИБРОКОМ». – 2012. – 248 с.

6. Нургалеева Т. Г. Аббревиация как средство экспрессивного словообразования [Электронный ресурс] : дис. … канд. филол. наук, спец. 10.02.04 «Германские языки» / Т. Г. Нургалеева. – Москва, 2010. – 240 с. – Режим доступа : http://www.dissercat.com/content/abbreviatsiya-kak-sredstvo-ekspressivnogo-slovoobrazovaniya

7. Jenkins K. Deciphering Journal Abbreviations with JAbbr // Code4Lib Journal. – 2009. – № 7. – URL : https://journal.code4lib.org/articles/1758

8. Mikolov T. Efficient Estimation of Word Representations in Vector Space [Electronic resourse] / T. Mikolov, K. Chen, G. Corrado, J. Dean // arXiv.org. – 2013. – URL : http://arxiv.org/pdf/1301.3781v3.pdf

9. Mikolov T. Distributed Representations of Words and Phrases and their Compositionality / T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean // Advances in Neural Information Processing Systems. – 2013. – P. 3111–3119.

10. Blei D. M. Latent Dirichlet Allocation / D. M. Blei, A. Y. Ng, M. I. Jordan // Journal of Machine Learning Research. – 2003. – № 3. – P. 993–1022.

11. Heinrich G. Parameter estimation for text analysis. – 2004. – URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.216.695

12. Шилов И.М. Автоматическое выявление и расшифровка аббревиатур и сокращений в тексте / Шилов И.М. – СПбГУ, 2016.

13. Кунгурцев А. Б. Разработка информационной технологии выделения терминов из документов на естественном языке / А. Б. Кунгурцев, С. Л. Зиноватная, Я. В. Поточняк, М. А. Кутасевич // Восточно-Европейского журнала передовых технологий. – 2018. – № 6.

References

1. Izbachkov YU. S. Informacionnye sistemy : uchebnik dlya vuzov / YU. S Izbachkov, V. N. Petrov. – Piter, 2011. – 544 s.

2. Kungurcev A. B. Metod avtomatizirovannogo postroeniya tolkovogo slovarya predmetnoj oblasti [EHlektronnyj resurs] / A. B. Kungurcev, YA. V. Potochnyak, D. A. Silyaev // Tekhnologicheskij audit i rezervy proizvodstva – 2015. – № 2/2(22). – S. 58 – 63. – Rezhim dostupa : http://nbuv.gov.ua/UJRN/Tatrv_2015_2%282%29__12.

3. Kunhurtsev O. B. Pobudova slovnyka predmetnoi oblasti na osnovi avtomatyzovanoho analizu tekstiv ukrainskoiu movoiu / O. B. Kunhurtsev, S. V. Kovalchuk, Ya. V. Potochniak, M. V. Shyrokostup // Tekhnichni nauky ta tekhnolohii – 2016. – № 3 (5). – S. 164–174.

4. Kungurcev A.B. Uchet mezhfrazovyh svyazej pri avtomatizirovannom postroenii tolkovogo slovarya predmetnoj oblasti / A. B. Kungurcev, A. I. Gavrilova, A. S. Leongard, YA. V. Potochnyak // Informatika i matematicheskie metody v modelirovanii. – 2016. – № 2. – S. 173–183.

5. Superanskaya A. V. Obshchaya terminologiya: Voprosy teorii. Abbreviaciya v terminologii / A. V. Superanskaya, N. V. Podol'skaya, N. V. Vasil'eva. – Izd. 6-e. – M. : Knizhnyj dom «LIBROKOM». – 2012. – 248 s.

6. Nurgaleeva T. G. Abbreviaciya kak sredstvo ehkspressivnogo slovoobrazovaniya [EHlektronnyj resurs] : dis. … kand. filol. nauk, spec. 10.02.04 «Germanskie yazyki» / T. G. Nurgaleeva. – Moskva, 2010. – 240 s. – Rezhim dostupa : http://www.dissercat.com/content/abbreviatsiya-kak-sredstvo-ekspressivnogo-slovoobrazovaniya

7. Jenkins K. Deciphering Journal Abbreviations with JAbbr // Code4Lib Journal. – 2009. – № 7. – URL : https://journal.code4lib.org/articles/1758

8. Mikolov T. Efficient Estimation of Word Representations in Vector Space [Electronic resourse] / T. Mikolov, K. Chen, G. Corrado, J. Dean // arXiv.org. – 2013. – URL : http://arxiv.org/pdf/1301.3781v3.pdf

9. Mikolov T. Distributed Representations of Words and Phrases and their Compositionality / T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean // Advances in Neural Information Processing Systems. – 2013. – P. 3111–3119.

10. Blei D. M. Latent Dirichlet Allocation / D. M. Blei, A. Y. Ng, M. I. Jordan // Journal of Machine Learning Research. – 2003. – № 3. – P. 993–1022.

11. Heinrich G. Parameter estimation for text analysis. – 2004. – URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.216.695

12. SHilov I.M. Avtomaticheskoe vyyavlenie i rasshifrovka abbreviatur i sokrashchenij v tekste / SHilov I.M. – SPbGU, 2016.

13. Kungurcev A. B. Razrabotka informacionnoj tekhnologii vydeleniya terminov iz dokumentov na estestvennom yazyke / A. B. Kungurcev, S. L. Zinovatnaya, YA. V. Potochnyak, M. A. Kutasevich // Vostochno-Evropejskogo zhurnala peredovyh tekhnologij. – 2018. – № 6.

Рецензія/Peer review : 15.11.2018 р. Надрукована/Printed :19.12.2018 р.
Рецензент: д.т.н., проф. Вичужанін В.В.