

СИСТЕМА АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕМАТИЧЕСКИ ОРИЕНТИРОВАННЫХ ТЕКСТОВ С ТЕРМИНОЛОГИЧЕСКИМ СЛОВАРЕМ В ФОРМАТЕ РЕГУЛЯРНЫХ ВЫРАЖЕНИЙ¹

В.С. Суховеров

Разработана система автоматической обработки текстов, оценивающая тематику текста по его терминологии, которая задается словарем терминов. Предложено и обосновано применение регулярных выражений в предметно ориентированных словарях, используемых в программах анализа текста на естественных языках. Отмечена и описана взаимосвязь регулярных выражений и конечных автоматов через регулярные множества. Предложена количественная оценка тематической направленности исследуемого текста — профиль документа, вычисляемый по результатам поиска терминов. Система реализована в виде программного комплекса с вариантом словаря для выбранной предметной области — теории управления и ее приложений. Система протестирована на архиве журнала «Автоматика и телемеханика». Получены профили тематической направленности статей, взятых из различных рубрик журнала. Указаны возможности развития системы.

Ключевые слова: термин, словарь предметной области, регулярное выражение, конечный автомат, профиль документа, программный комплекс.

ВВЕДЕНИЕ

Потребность в регулярных выражениях [1] возникла при создании системы автоматического анализа тематики русскоязычных текстов, принадлежащих определенной предметной области. Задача определения тематики научного текста, его принадлежности к тем или иным разделам определенной прикладной области вполне актуальна и востребована. Она зачастую является составной частью таких информационных задач, как классификация и рубрикация, а также задач управления, связанных с принятием решений: подбор научных сотрудников для выполнения проекта, распределение тематических заданий в редакциях научных журналов и др. Анализируя тексты научных публикаций специалистов, можно оценить их

компетентность или их соответствие компетенциям, предъявляемым для участия в определенном проекте.

Предлагаемый подход ограничивается тематической оценкой текста и не анализирует качество документа (новизну, актуальность, грамотность изложения и т. д.), поэтому предполагает участие эксперта на стадии принятия решения. Эти вопросы обсуждались в предыдущей работе [2].

В настоящей работе излагается подход к построению механизма автоматической обработки текста и основных его составляющих: программы поиска и словаря предметной области, а также подход к организации их совместной работы. Определение тематики научного текста можно представить как совокупность двух задач:

- отыскание в тексте всех возможных словоформ терминов из словаря предметной области;
- тематическая классификация найденных терминов и количественная оценка тематических составляющих.

¹ Работа выполнена при частичной финансовой поддержке РФФИ (проект 18-07-01044а).

Основное требование к программе поиска терминов в тексте документа — это точность и полнота результата. Программа должна найти в тексте все возможные словоформы терминов из словаря предметной области и только их, если таковые содержатся в документе. Известно множество готовых программных средств для анализа текстов на естественных языках: стеммер Портера, АОТ (автоматическая обработка текста) [3], *mystem* [4] (морфологический анализатор русскоязычных текстов), *ruMorphy* [5] (на основе словарей и идей АОТ) и др. Стеммер Портера находит слово по его неизменяемой части и является наиболее простым и широко применяемым средством, как правило, в различных комбинациях с дополнительными алгоритмами морфологического, синтаксического и семантического анализа. Из числа полнофункциональных программных систем отметим АОТ, которая выполняет полный анализ текста, охватывая все уровни представления: графематический, морфологический, синтаксический и семантический. Но сложность и объемность этой системы препятствует ее адаптации для решения других задач. Перечисленные примеры программных средств в настоящей работе не применялись.

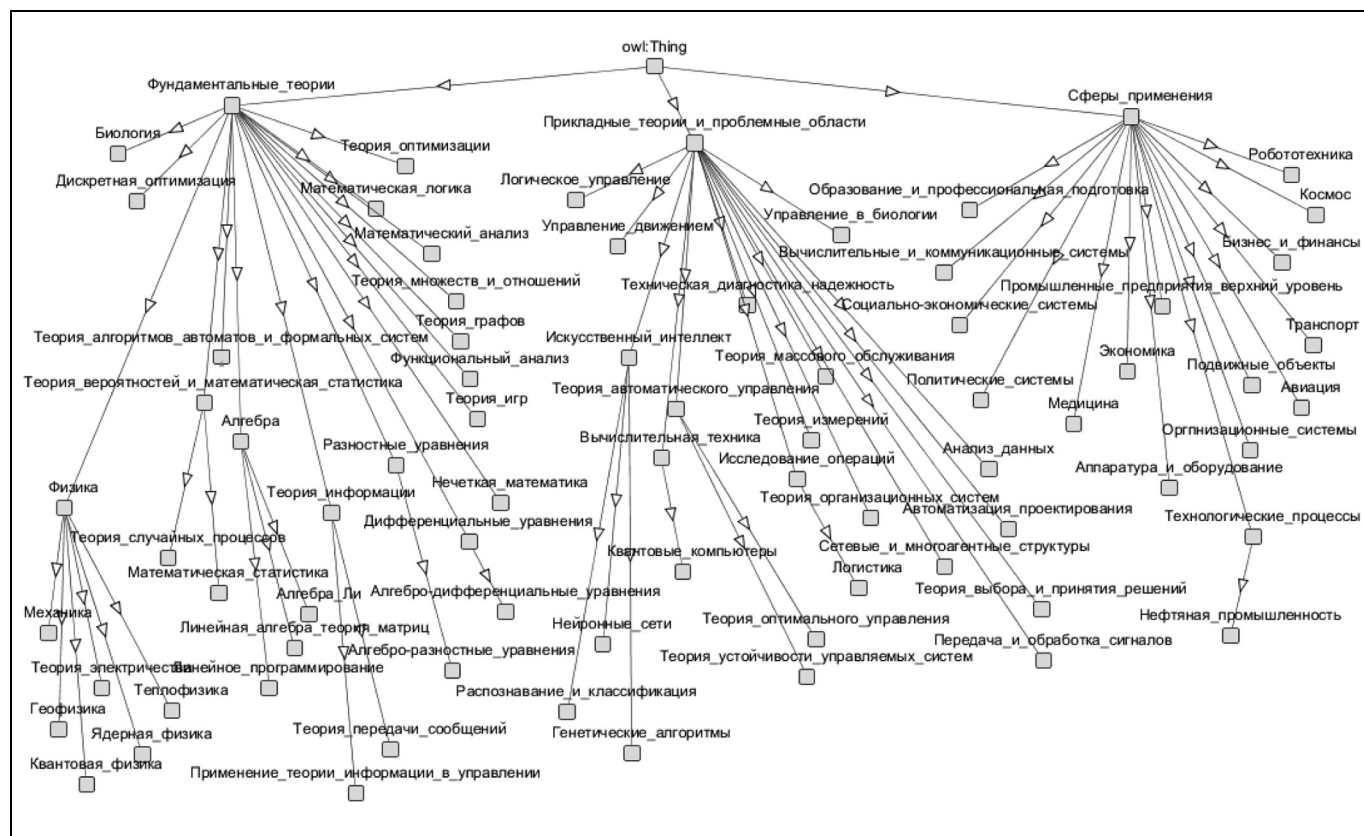
1. О СИСТЕМЕ В ЦЕЛОМ

В системе можно выделить две основные функциональные составляющие: словарь предметной области и программу поиска терминов. Автономный словарь представляет собой книгу Excel. Тематика словаря должна соответствовать предметной области, в которой работает система. Замена словаря предметной области не требует изменения программного обеспечения. Словарь базируется на онтологии [6] предметной области. В данном случае был составлен и использовался словарь терминов теории управления. Онтологический подход к описанию предметной области подробно изложен в работе [2]. В словаре структура связей дерева онтологии преобразована в рубрикатор. Терминологическая информация словаря представлена в формате регулярных выражений. Регулярные выражения в предлагаемом подходе служат наиболее подходящей моделью для записи терминов словаря, так как удовлетворяют требованиям точного и полного описания терминов. Применение регулярных выражений существенно упростило задачу поиска терминов в тексте анализируемого документа, поскольку ряд языков программирования содержит библиотеки для работы с регулярными выражениями. В окончательном виде словарь предметной области представляет собой книгу Excel, содержащую табличные представления рубрикатора и терминологии в формате ре-

гулярных выражений. Программа поиска кроме стандартных библиотек Python, содержит ряд сторонних программных пакетов, предназначенных для работы с регулярными выражениями. Программа поиска встроена в программный комплекс [7], который помимо нее содержит программный интерфейс для работы со словарем терминов, программу для расчета профиля документа или совокупности документов и экранный интерфейс пользователя.

2. СЛОВАРЬ ПРЕДМЕТНОЙ ОБЛАСТИ

Термины предметной области отбирались из материалов архива журнала «Автоматика и телемеханика». Структура онтологического дерева (рис. 1) выстраивалась параллельно с распределением отобранных терминов по темам. В результате определилось дерево онтологии, которое имеет три выраженные ветви: «фундаментальные теории», «прикладные теории и проблемные области», «область приложений». Термины в основном сосредоточены в листьях дерева графа предметной области. При таком наполнении онтологии мы получаем наиболее дифференцированную информацию о тематической направленности текста. Дополнительно можно отметить, что при создании онтологии использовались далеко не все возможности пакета Protégé-OWL [8, 9], относящиеся к построению аксиоматики, определению свойств и взаимоотношений классов и экземпляров и др. Это объясняется тем, что в рассматриваемом случае структура онтологии достаточно простая. Связи классов онтологии «тема — подтема» интерпретируются как «класс — подкласс» с отношением «is — a». При этом классы онтологии содержат экземпляры одного типа (термины), которые обладают лишь одним признаком — принадлежностью теме. Его значение изменяется в пределах от 0 до 1 и зависит от числа тем, которым принадлежит термин. Термины могут быть однословными и многословными. При наполнении словаря важно было не включать в качестве самостоятельных однословных терминов широко распространенные в научной и научно-технической литературе слова, такие как «алгоритм», «задача», «цикл». Они вошли в состав многословных терминов, специфических для определенных тем: «генетический алгоритм», «задача коммивояжера», «гамильтонов цикл». Для удобства применения в системе поиска терминов информация онтологии предметной области преобразована в формат книги Excel, которая содержит таблицу-рубрикатор тем онтологии и таблицу-словарь терминов предметной области, а также служит для отображения результатов обра-



ботки текстов, получаемых при работе системы, например, графических построений по результатам тестирования системы. Текущая версия словаря онтологии не обеспечивает полноты терминов и списка тем: они будут постоянно расширяться в ходе эксплуатации и по мере привлечения специалистов в данной области знаний. Естественно, научные направления непрерывно развиваются, появляются новые теории и области приложений. Это приводит к расширению и уточнению терминологии. Поэтому и словарь, и онтология должны быть постоянно открытыми для расширения и модификации.

В электронных общелексических словарях с большими объемами информации первоочередная задача состоит в упорядочении информации и применении алгоритмов поиска с наименьшими ожидаемыми либо гарантированными временными затратами, например, применение модели на конечных автоматах с применением алгоритма Томпсона [10]. Специфика рассматриваемой задачи состоит в том, что словарь терминов предметной области несравнимо меньше по объему общелексического словаря. Даже при включении в него, кроме однословных и многословных терминов,

еще и характерных контекстных оборотов объем словаря вряд ли превысит 10^4 единиц. Поэтому в данном случае при разработке программы поиска не было жестких требований ни к скорости поиска, ни к объему оперативной памяти. Отметим, что задача объединения терминологической части словаря в единую морфосинтаксическую структуру не ставилась. Важно было найти такую модель представления терминов, которая бы явилась самодостаточной, т. е. не требующей дополнительных средств анализа текста при определении терминов предметной области. Оказалось, что это можно сделать с помощью регулярных выражений, причем для каждого термина (однословного или многословного) все его возможные словоформы можно объединить в одно регулярное выражение.

3. РЕГУЛЯРНЫЕ ВЫРАЖЕНИЯ В СИСТЕМЕ ПОИСКА ТЕРМИНОВ

Исходя из строковых представлений, регулярные выражения можно рассматривать как искусственный формальный язык, в котором существует два типа символов: литералы и метасимволы (или управляющие символы). Из литералов фор-

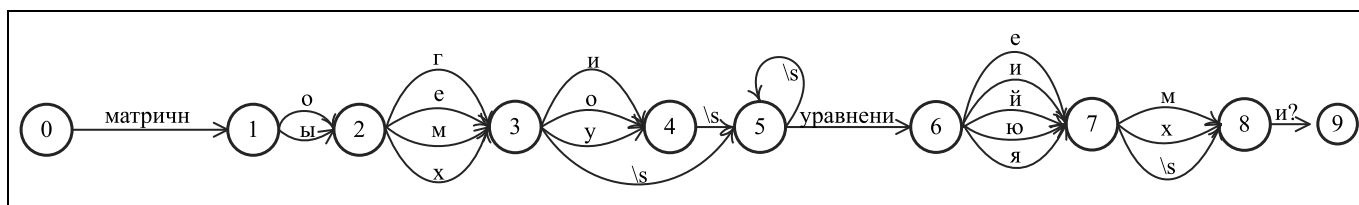


Рис. 2. Регулярное выражение в интерпретации конечного автомата со сжатием ключей на дугах $0 \rightarrow 1$ и $5 \rightarrow 6$

мируются слова текста, а метасимволы выполняют функции грамматических элементов. Набор метасимволов в регулярных выражениях обладает большими выразительными возможностями. Слова, объединенные по определенным правилам с грамматическими элементами, образуют структурированные конструкции, способные компактно выразить все возможные словоформы словосочетаний, включая однословные и многословные термины, а также характерные для данной предметной области контекстные обороты. Рассмотрим пример двусловного термина.

Пример: матричное уравнение. Регулярное выражение, объединяющее все словоформы этого термина, имеет вид:

$((\text{матричн})([\text{оы}][\text{гемх}][\text{иоу}\backslash\text{s}])\backslash\{1,2\})(\text{уравнени})$
 $([\text{ейиюя}][\text{мх}\backslash\text{s}]\text{и}?) \blacklozenge$

Из примера видно, что регулярные выражения представляют собой строки, структурированные с помощью управляющих символов. Автором регулярных выражений является математик Стивен Коул Клини (*Stephen Cole Kleene*). Он дал формальное описание этих моделей при помощи алгебры регулярных множеств. Теорема Клини доказывает, что для каждого регулярного выражения R может быть построен конечный автомат (возможно, недетерминированный), распознающий язык, задаваемый выражением R . Другими словами: классы регулярных множеств и автоматных языков совпадают. В рассматриваемой ситуации каждое регулярное выражение, представляющее, в общем случае, многословный термин, является элементом регулярного множества (словаря терминов предметной области), построенного на алфавите языка регулярных выражений, и в этом случае в программе поиска можно ограничиться применением парсера. При этом каждую конструкцию термина в виде регулярного выражения можно представить и как конечный автомат со сжатием. В данном случае каждый такой конечный автомат, представляющий термин, имеет один вход и один выход.

Для работы с регулярными выражениями разработаны не только алгоритмы, но и, в ряде языков программирования, библиотеки методов для выполнения различных операций и, в первую оче-

редь, поиска. Для удобства применения регулярных выражений в программе поиска термины онтологии были представлены в словаре двояко: а) в лемматизированной или «нормальной» форме (существительные — именительный падеж, единственное число; прилагательные — именительный падеж, единственное число, мужской род; глаголы, причастия, деепричастия — глагол в инфинитиве); б) в виде регулярных выражений. Между представлениями а) и б) установлено взаимно-однозначное соответствие. Такой подход обеспечивает применение в программе поиска регулярных выражений, поскольку лемматизированная форма не охватывает все словоформы терминов и ее применение требует дополнительного морфологического и синтаксического разбора текстовых фрагментов. Если возникнет ситуация, предполагающая серьезное увеличение объема словаря, то применение регулярных выражений обладает определенной перспективой, а именно, возможностью перехода на более быстрые алгоритмы поиска, построенные на конечных автоматах (рис. 2), например, алгоритм Томпсона. Преимущества этого алгоритма по сравнению с бэктрекингом изложены в работе [11].

4. ОЦЕНКА ТЕМАТИЧЕСКОЙ НАПРАВЛЕННОСТИ ТЕКСТА

Для количественной оценки релевантности документа темам предметной области в систему введен профиль документа [2], который вычисляется по результатам поиска терминов. Пусть в документе D найдено K_D вхождений терминов, принадлежащих онтологии, содержащей N тем; темы и термины пронумерованы. Определим значимость термина для темы. Под значимостью i -го термина для j -й темы будем понимать величину s_j^i , такую, что:

$s_j^i = 1/c_j$, если i -й термин принадлежит j -й теме и c_i — число тем, которым принадлежит i -й термин;

$s_j^i = 0$, если i -й термин не принадлежит j -й теме.

Таким образом, принадлежность термина нескольким темам уменьшает вероятность его при-

надлежности каждой из них. Значимость всех найденных терминов K_D для j -й темы:

$$S_j = \sum_{i=1}^{K_D} s_j^i.$$

Суммарная значимость всех найденных терминов документа D по всем темам онтологии:

$$S_D = \sum_{j=1}^N S_j.$$

Релевантность $R_j(D)$ документа D j -й теме определяется как сумма значимостей всех терминов j -й темы, нормированная по S_D :

$$R_j(D) = S_j / S_D.$$

При таком нормировании сумма релевантностей документа всегда равна 1.

В результате, профиль документа можно определить как вектор релевантностей в пространстве тем онтологии: $\overline{P(D)} = (R_1(D), R_2(D), \dots, R_N(D))$.

5. ПРОГРАММНЫЙ КОМПЛЕКС. ТЕСТИРОВАНИЕ И ПРИМЕНЕНИЕ

Разработанный программный комплекс [12] предназначен для автоматического определения тематической принадлежности документов к определенной предметной области. Он выполняет предварительную обработку входного документа, а затем производит поиск в тексте документа терминов, однословных и многословных, принадлежащих словарю предметной области, строит профиль документа и выводит результат на экран, в файл или на печать. Документы должны быть представлены файлами в форматах TXT или PDF.

Программный комплекс написан на языке *Python* [13]. Для работы с регулярными выражениями в арсенале языка *Python* имеется мощная библиотека, которая позволяет компилировать регулярные выражения в объекты шаблонов и содержит методы для различных операций с ними, таких как поиск вхождения шаблона или выполнение замены строки. В разработке использовались и дополнительные пакеты программных модулей, в частности, пакет *pdfminer* для работы с файлами формата PDF.

Программный комплекс содержит: основную программу, выполняющую в автоматическом режиме поиск терминов в тексте документа и вычисление профиля документа (рис. 3); экранный интерфейс пользователя; словарь предметной области (книгу Excel) и программный интерфейс для работы с ним.

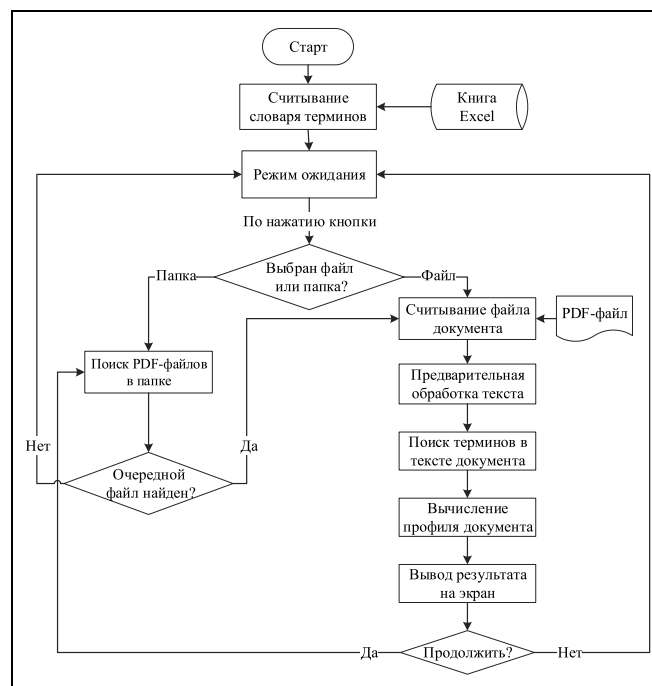


Рис. 3. Структурная схема программы автоматического определения профиля документа

Главное окно экранного интерфейса содержит кнопки управления обработкой файлов и панель для вывода результатов. Кроме главного окна, в состав интерфейса входит несколько диалоговых окон, обеспечивающих выбор режимов работы программного комплекса.

Тестирование проводилось на файлах текстового формата. Протестированы две рубрики из архива журнала «Автоматика и телемеханика» — «Системы массового обслуживания» и «Теория линейных систем». Результаты тестирования и их графическая интерпретация приведены в работе [10]. Еще один вариант тестирования на материалах архива журнала «Автоматика и телемеханика» проведен для членов редколлегии этого журнала. Тест заключался в сопоставлении сведений о научных интересах, представленных тестируемыми членами редколлегии, с профилем научных интересов, полученным системой автоматически по совокупности их публикаций в журнале «Автоматика и телемеханика».

Один из примеров.

Представлено членом редколлегии:

стохастическое программирование, стохастическое управление, стохастическое оценивание, минимаксная оптимизация, теория нечетких множеств, дискретная оптимизация, двухуровневое программирование, приложения теории оптимизации к аэрокосмическим системам и к финансовой математике, построение системы дистанционного обучения.

Получено с помощью системы автоматической обработки:

линейная алгебра, теория матриц 0,3 %;
линейное программирование 3,6 %;
дифференциальные уравнения 0,8 %;
дискретная оптимизация 2,9 %;
математический анализ 0,6 %;
функциональный анализ 0,9 %;
теория алгоритмов 0,3 %;
теория вероятностей и математическая статистика 29,7 %;
математическая статистика 1,6 %;
теория графов 1,0 %;
теория игр 12,3 %;
теория информации 0,1 %;
теория множеств и отношений 2,5 %;
теория оптимизации 1,3 %;
механика 2,5 %;
биология 0,1 %;
вычислительная техника и программирование 0,2 %;
искусственный интеллект 1,1 %;
генетические алгоритмы 0,0 %;
распознавание и классификация 0,1 %;
исследование операций 2,7 %;
передача и обработка сигналов 0,1 %;
теория автоматического управления 13,0 %;
теория устойчивости управляемых систем 0,1 %;
теория оптимального управления 11,5 %;
теория выбора и принятия решений 1,7 %;
теория массового обслуживания 1,4 %;
техническая диагностика, надежность 1,8 %;
управление движением 0,2 %;
авиация 0,1 %;
бизнес и финансы 4,6 %;
экономика 0,5 %.

Всего найдено терминов в файлах: 1492.

Соответствие полученных оценок научным интересам членов редколлегии следует воспринимать как ориентировочное и приблизительное по нескольким причинам. Одна из них — неполные списки публикаций, поскольку авторы публикуются, как правило, не в одном периодическом издании. Существуют и другие причины — это неполнота онтологии и неравномерное наполнение классов онтологии терминами. И если неполнота онтологии устраняется благодаря пополнению словаря экспертами, то неравномерность наполнения классов — это естественное явление, обусловленное тем, что различным тематикам соответствует разное количество характерных терминов. Это обстоятельство будет учтено в дальнейшем, оно предполагает ввод дополнительных весовых коэффициентов для значений значимости терминов при вычислении профиля документа.

ЗАКЛЮЧЕНИЕ

Представленная система показала целесообразность применения регулярных выражений в качестве терминов, составляющих словарь предметной

области. Как было отмечено ранее, замена словаря предметной области не требует изменения программного обеспечения. Это означает, что система допускает возможность применения словарей на иностранных языках. В настоящее время система развивается. Ряд дальнейших усовершенствований системы был перечислен в работе [2]. Разрабатывается сетевой вариант системы для работы с сайтом ИПУ РАН, где размещены электронные копии научных публикаций (монографий, статей, докладов и т. д.) сотрудников и аннотации этих публикаций. Реализована и опробована предварительная версия сетевого варианта системы, работающая с файлами аннотаций докладов и публикаций.

ЛИТЕРАТУРА

1. Фридл Дж. Регулярные выражения. Пер. с англ. — 3-е изд. — СПб.: Символ-Плюс, 2008.
2. Кузнецов О.П., Суховеров В.С. Онтологический подход к оценке тематики научного текста // Онтология проектирования. — 2016. — Т. 6, № 1 (19) — С. 55–66.
3. Автоматическая обработка текста / Технологии. — URL: <http://aot.ru/technology.html> (дата обращения 20.09.2018).
4. Mystem. — URL: <https://nlp.ru/Mystem> (дата обращения: 20.09.2018).
5. Морфологический анализатор pymorphy2. — URL: <https://pymorphy2.readthedocs.io/en/latest/> (дата обращения: 20.09.2018).
6. Кузнецов О.П., Суховеров В.С., Шипилина Л.Б. Онтология как систематизация научных знаний: структура, семантика, задачи. — М.: ИПУ РАН, 2010. — С. 126–138.
7. Суховеров В.С. Оценка тематики научного текста на основе онтологического описания предметной области, практика применения // Сб. науч. тр. XX юбилейной Российской науч. конф. «Инжиниринг предприятий и управление знаниями» (ИП & УЗ—2017, Москва). — М.: РЭУ им. Г.В. Плеханова, 2017. — Т. 1. — С. 187–193.
8. Smith M.K., Welty C., McGuinness D.L. OWL Web Ontology Language Guide, 2004. — URL: https://www.researchgate.net/publication/228057927_OWL_Web_Ontology_Language_Guide (дата обращения: 24.12.2018).
9. Horridge M. A Practical Guide to Building OWL Ontologies Using Protégé 4 and CO-ODE Tools. — The University of Manchester, 2011.
10. Thompson K. Programming Techniques: Regular expression search algorithm // Communications of the ACM. — 1968. — Vol. 11, iss. 6. — P. 419–422.
11. Cox R. Regular Expression Matching Can Be Simple And Fast (but is slow in Java, Perl, PHP, Python, Ruby, ...). — January 2007. — URL: <https://swtch.com/~rsc/regexp/regexp1.html> (дата обращения: 20.09.2018).
12. Суховеров В.С. Программный комплекс для автоматического определения профиля документов, относящихся к определенной предметной области: Свид. о гос. регистрации программы для ЭВМ № 2018615473 РФ. Зарег. 08.05.2018.
13. Бизли D. Python. Подробный справочник: пер. с англ. — СПб.: Символ-Плюс, 2010.

Статья представлена к публикации членом редколлегии В.Г. Лебедевым.

Суховеров Виктор Степанович — канд. техн. наук, науч. сотрудник, Институт проблем управления им. В.А. Трапезникова РАН, г. Москва, ✉ vicst09@yandex.ru.

Поступила в редакцию 27.09.2018, после доработки 22.10.2018.

Принята к публикации 12.12.2018.