УДК 654.07 81’33
https://doi.org/10.38086/2522-9613-2019-2-46-52

Интеллектуальный анализ текстовой информации в специализированных областях в системе электронного правительства

Т. И. Макаревич, м. ф. н., старший преподаватель кафедры английского языка гуманитарных специальностей факультета международных отношений[1], магистрант 1 курса специальности «Электронное правительство» факультета инновационной подготовки[2]

E-mail: t_makarevich@mail.ru

ORCID ID: 0000-0002-3720-2373

[1] Белорусский государственный университет, ул. Ленинградская, д. 20, 220030, г. Минск, Республика Беларусь

[2] Академия управления при Президенте Республики Беларусь, ул. Московская, д. 17, 220007, г. Минск, Республика Беларусь

Аннотация. Настоящая статья посвящена изучению применения технологии text mining в научных исследованиях как одного из методов интеллектуального анализа текстовой информации в специализированных областях системы электронного правительства. Значимость работы объясняется тем, что в настоящее время в Республике Беларусь не существует исследований, аналогичных проведенному. Продемонстрировано применение программного пакета Rapid Miner и языка R как сред для глубинного анализа текста. Оптимальной формой изучения предметных онтологий признано так называемое концептуальное индексирование. Обозначены оптимальные подходы к его рассмотрению: формальный и лингвистический. Выявлены проблемы избыточности и многозначности слов. Разработка данной проблематики нацелена на согласование разрозненности русскоязычных и иноязычных терминологических систем специализированных онтологий на основе технологий искусственного интеллекта.

Ключевые слова: терминологическая система, специализированные терминологические словари, информационно-поисковый тезаурус, онтология, предметная область, обработка текстовой информации, частотный анализ, глубинный анализ текста, язык R, Rapid Miner, электронное правительство

Для цитирования: Макаревич, Т. И. Интеллектуальный анализ текстовой информации в специализированных областях в системе электронного правительства / Т. И. Макаревич // Цифровая трансформация. – 2019. – № 2 (7). – С. 46–52. https://doi.org/10.38086/2522-9613-2019-2-46-52

Intellectual Analysis of Textual Information in Domain Fields in the System of e-Government

T. I. Makarevich, Master of Philological sciences, Senior Lecturer of the Department of English for Humanities, Faculty of International Relations[1], 1st year postgraduate student, specialty “e-Government”[2]

E-mail: t_makarevich@mail.ru

ORCID ID: 0000-0002-3720-2373

[1]Belarusian State University, 20 Leningradskaya Str., 220030 Minsk, Republic of Belarus

[2]Academy of Public Administration under the aegis of the President of the Republic of Belarus, 17 Moskovskaya Str., 220007 Minsk, Republic of Belarus

Abstract. The given paper considers application of data mining technology in scientific research as one of intellectual analysis methods in the domain field of e-Government. The topicality of the issue is stipulated by the current absence of the researches of the kind in the Republic of Belarus. The paper illustrates how the programme package Rapid Miner and the language R have been applied in text mining. Concept indexing has been admitted as the most resultative

form of analyzing domain field ontologies. Formal and linguistic approaches are found most effective in analyzing domain field ontologies. The paper identifies the problems of word redundancy and word polysemy. The prognosis for the further research investigation is in interconnectivity of specialized ontologies studying heterogeneous terms on the basis of artificial intelligence (AI).
Key words: terminilogical system, specialized dictionaries, information retrieval thesaurus, ontology, domain area, text information processing, analysis in the frequency domain, text mining, the computer language R, Rapid Miner, e-Government

For citation: Makarevich T. I. Intellectual Analysis of Textual Information in Domain Fields in the System of e-Government. *Cifrovaja transformacija* [Digital transformation], 2019, 2 (7), pp. 46–52 (in Russian). https://doi.org/10.38086/2522-9613-2019-2-46-52

Введение. Одна из основных проблем в интеллектуальном анализе текста заключается в необходимости постоянного обновления терминологических словарей естественных языков в сфере информационно-коммуникационных технологий (ИКТ), а также в широте предметных областей их применения, которая обусловлена высокой интенсивностью развития систем искусственного интеллекта (AI systems), и, в частности, машинного обучения.

Актуальность темы обусловлена высокой динамичностью развития систем искусственного интеллекта и процессов в сфере представления знаний в области ИКТ; проблемой взаимосвязи терминов при создании тезаурусных отношений; разработкой терминологических систем, основанных на онтологиях и тезаурусах; удовлетворительным решением задач с их помощью с последующим представлением решений в системе электронного правительства (ЭП).

Объектом исследования является лингвистическое обеспечение информационных систем, предметом – методика проведения частотного анализа специализированных терминов.

Цель работы заключается в проведении интеллектуального анализа текстовой информации в специализированных областях с созданием рабочего макета лингвистического обеспечения информационных систем для согласования данных русскоязычной и иноязычных терминологических систем предметных областей в системе ЭП.

В задачи работы входит: изучение методики актуальной обработки текстовой информации; исследование возможности формирования и корректировки онтологических моделей на основе интеллектуальной обработки текстовой информации (text mining); проведение анализа использования программного обеспечения языка R и программного пакета Rapid Miner как оптимальных программных средств для создания информационной системы интеллектуальной обработки текстовой информации. Для решения поставленных задач мы используем метод глубинного анализа текстовой информации.

В работе рассматриваются теоретические и прикладные вопросы лингвистического обеспечения информационных систем, раскрывается сущность терминологического моделирования специализированных областей для формирования лингвистического обеспечения информационных систем. Аналитически рассматриваются онтологические модели как способ описания предметных областей.

Проводилось проектное исследование технологии text mining для анализа текстов на предмет поиска в них заданной терминологии определенных предметных областей. В работе рассматривается технология data mining как процесс выделения и сортировки неструктурированных данных с последующим представлением для дальнейшего прикладного использования, описывается практическое применение технологии Rapid Miner для глубинного анализа текста.

Практическая значимость результатов работы заключается в возможности их применения при обработке естественного языка в режиме online, в процессах информационного поиска, категоризации текстов, автоматической обработке больших текстовых коллекций (например, в корпусной лингвистике), появлении возможности формирования и корректировки онтологических моделей с применением text mining.

Основоная часть. Наблюдаемая сегодня необходимость обработки неструктурированной текстовой информации, широкая предметная область её применения [1], повышение эффективности и качества уже существующих методов обработки текстов – всё это продиктовано современными потребностями работы с большими объемами электронных документов в системе ЭП. На смену понятию big data (неопределённо структурированные данные большого объёма) [2] приходит новый концепт – smart data («умные данные», результаты, извлеченные за счет обра-

ботки «больших данных»; данные, полученные в результате интеллектуального анализа данных) [2]. Сегодня его рассматривают как целое направление, поскольку важным становится не столько сам объем данных, как то, каким образом и в какой области они могут быть использованы.

В зависимости от поставленных целей перед процессом обработки неструктурированной текстовой информации исследователи занимаются такими задачами, как автоматическое аннотирование документов, поиск ответов на вопросы, непосредственный поиск информации, фильтрация, рубрикация, кластеризация документов и групп документов, поиск схожих документов, их сегментирование и смежные вопросы.

Наше исследование нацелено на создание механизма отслеживания способности предоставления системой электронного правительства возможности взаимодействия специализированных онтологий различных предметных областей, а также на сопоставление русскоязычных и англоязычных специализированных онтологий, выявление в них проблемных зон, лингвистических лакун, нестыковок терминологии с целью последующего решения данных вопросов. Поскольку ЭП призвано осуществлять взаимодействие организаций всех типов и уровней, то в систему ЭП, наряду с государственными, включены и научные организации. Вход в систему специализированных онтологий широких предметных областей должен контролироваться системой ЭП, которая должна также быть способна обеспечить организациям и специалистам различных предметных областей непрерывный доступ к специализированным онтологиям, иметь способность к согласованию терминосистем данных онтологий.

Для эффективности работы ЭП в современных информационно-поисковых и информационно-аналитических системах ведется работа с текстовой информацией в неограниченных специализированных областях, включающих в себя различные классы сущностей и предполагающих возможность вхождения в неограниченное многообразие отношений между собой [1]. При этом существуют такие нерешенные вопросы, как, например, нехватка лингвистических и онтологических знаний (знаний о мире), использующихся в приложениях информационного поиска и автоматической обработке текстов, что может привести к нерелевантному поиску в случае если формулировка запроса отличается от способа описания релевантной ситуации в текстовом документе. В системе ЭП эта проблема может усугубиться при обработке так называемых длинных запросов, в поиске ответа на вопрос в вопросно-ответных системах.

На данном этапе развития системы ЭП внедрение дополнительных объемов знаний о языке и мире в современные методы автоматической обработки текстов представляет собой сложную задачу. Это обусловлено тем, что, во-первых, подобным знаниям следует давать описание в специально создаваемых ресурсах, таких как тезаурусы или онтологии, содержащие описания десятков тысяч слов и словосочетаний с возможностью иметь операции по логическому выводу. Во-вторых, возникает необходимость автоматизации решения проблемы многозначности слов через выбор наиболее подходящего значения. Также стоит учитывать, что развитие актуальных комбинированных методов (знания плюс последние разработки в обработке цифровой информации) происходит в условиях, когда ведение любых ресурсов отстает от развития предметной области.

Применение технологии data mining, которая активно используется для обнаружения необходимых знаний в базах данных (так называемом KDD – Knowledge Discovery in Databases [3]), нужно для эффективного функционирования системы ЭП с целью научного поиска. Задача text mining заключена в процессе построения модели, хорошо описывающей закономерности, которые порождают данные. При анализе текстовой информации применяются различные методы исследования data mining: «деревья решений», построение нейронных сетей, применение методов ограниченного перебора, кластерные модели, генетические алгоритмы, комбинированные методы, эволюционное программирование. Процесс data mining является итеративным. Это значит, что при решении какой-то конкретной задачи провоцируются новые, которые нужно решать, пока не будет достигнут удовлетворительный результат.

В настоящее время технология data mining используется в тех сферах деятельности человека, где есть накопление ретроспективных данных: наука, индивидуальное предпринимательство, а также веб-направление [3, с. 74]. Так, в нашей работе мы не будем перечислять все сферы применения интеллектуального анализа данных, нас интересует вопрос использования систем интеллектуального анализа данных как инструмента в проведении уникальных исследований.

С целью точности формулировки понятий приведем определение технологии data mining

одного из основателей этого направления Г. Пиатецкого-Шапиро [3], который обозначает ее как «процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности» [4].

Для проведения интеллектуального анализа текстовой информации при помощи компьютера мы применили программный пакет Rapid Miner как среду для глубинного анализа текста. Следует отметить, что на русском языке о text mining мало информации и на данный момент существует ограниченное количество пошаговых руководств в работе с той или иной программой, видеотренингов, чтобы можно было проанализировать специализированные тексты.

Text mining призван решить ряд определенных задач, первой из которых является классификация текста, основная идея которой заключается в том, что документы, принадлежащие к одной категории, содержат одинаковые слова и словосочетания [5]. Мы заранее создаем и разделяем текст на категории, к которым он относится, далее идет машинное обучение, чтобы мы могли понять, относятся ли эти тексты к той или иной группе.

Следующая по важности задача – классифицирование текстов с помощью машинного обучения, т. е. с обучающей выборкой, где можно использовать экспертный метод выделения признаков и составления правил. Если с машинным обучением процедура понятна, то при экспертном методе формируется словарь на основе набора терминов в предметной области и соотнесения между ними. Документ классифицируется к своей рубрике на базе частоты, с которой появляются выделенные в тексте термины.

Третья задача – кластеризация сложного текста. Типичный кластерный анализ состоит в объединении имеющих схожесть текстов в группы. Эти данные должны быть представлены в модели векторного пространства. Кластеризация может быть иерархической (похожа на метод интеллектуального анализа текста «дерево решений») и бинарной (характеризуется группировкой документальных кластеров по ссылкам подобия).

Четвёртая – это аннотирование текстов, который необходим для быстрого ознакомления с интересующей публикацией. Однако, благодаря возможностям text mining, анализ текста не ограничивается только этой функцией.

При помощи text mining в анализе текста мы также рассматриваем кластеризацию текста в программе Rapid Miner [5] и правило ассоциаций в этой среде. Для этого нужны исходные данные – тексты. Для репрезентативности выборки мы кластеризовали 185 английских текстов ИКТ-дискурса, дипломатического дискурса и дискурса международного права на примере валидного массива.

Основы обработки текста вытекают из объективности того факта, что именно в этом формате чаще всего предоставляется информация в Интернете. В отличие от технологии data mining, где осуществляется извлечение неочевидных данных, задачей обработки текста – text mining – является извлечение очевидныхданных . Например, при помощи данной технологии можно создать некоторый алгоритм [5], позволяющий извлечь большое количество информации с новостных сайтов, в нашем случае ИКТ-дискурса, дипломатического дискурса и дискурса международного права.

Первым этапом при обработке текстов в text mining является работа с коллекцией документов – это либо набор файлов в определенном формате, либо поток, который поступает на вход из вне. Следующий этап – декодирование (Def.) – перевод последовательности байт в последовательность символов. Затем нужно провести их распаковку (plain, .zip, .gz), и форматирование (csv, xml, json, doc). При прохождении этих трех этапов преобразования байтов в слова мы получаем текст в необходимом виде.

Этап токенизации заключается в разделении текста на токены, (последовательность символов/слов в памяти, понятную для машины, полученную не через наивный подход, используя n-граммы). Следующий шаг – это удаление стоп-слов, не несущих существенной информации для текста.

Нормализация как ещё один этап text mining является одним из самых сложных и состоит в приведении токенов к единому виду. Это необходимо для того, чтобы избавиться от поверхностной разницы в написании. На этом этапе важно сформулировать правила, по которым идет разбиение на токены.

Стемминг и лемматизация используются на предпоследнем этапе процесса text mining. Стемминг – это первый подход, который заключается в приведении грамматических форм слова, а также однокоренных слов к единой основе. Лемматизация – более сложный подход, в котором используется морфологический анализ с применением словарей и профессиональный подход к лингвистике. Она направлена на выделение лемм и оставление только тех слов, которые нужны [5, с. 24].

В программный пакет Rapid Miner включены подготовленные готовые процессы для работы с текстом. Вначале работы в программном продукте Rapid Miner необходимо создать новый процесс, в данном случае процессы для анализа текста уже созданы в самой программе. Их необходимо импортировать через file-import. Также нужны папка локального репозитория и файл process01.rmp, в которой находятся операторы для первого процесса, посвященного установлению ассоциативных связей. При этом каждый оператор имеет свои параметры.

Для первого процесса параметр директории текста указывает, откуда читать текстовые данные. Очень важным является вектор создания. Для него был выбран метод TF-IDE (term frequency-inverse document frequency, частота термина). Это метод взвешивания термина, обратная частота документа, инверсия той частоты, с которой конкретное слово встречается в коллекции документов [6]. Он дает наибольший вес тем терминам, которые встречаются наиболее часто в одном документе, но не много раз в других. Применение данного метода может дать слишком много слов, вплоть до 10 тысяч. Его недостатком является длительное время выполнения алгоритмов в случае слишком большого числа слов, что препятствует его использованию для последовательных этапов интеллектуального анализа.

Анализ текстов ИКТ-дискурса показал, что слова, которые появляются в менее чем 70% документов, сокращены. В поиске ассоциаций нас интересуют термины в контексте анализа текста, появляющиеся в 100% операций, т.е. документов в контексте проводимого интеллектуального анализа текстовой информации. Данные термины могут образовывать интересные часто встречающиеся наборы (т.е. список слов) и правила ассоциации для обеспечения действительного понимания. Поэтому целесообразно устанавливать параметр сокращения выше 100% включая слова, которые появляются в каждом документе.

Прежде всего, мы указали набор текстовых данных, который обрабатывался в process01 [5]. Process01 содержит 5 операторов. Первый процесс «documents from file» [5, с. 12] из оператора файлов выполняет обработку текста, которая включает подготовку текстовых данных для применения традиционных техник data mining. Процесс «documents» считывает данные с коллекции текстов и манипулирует этими данными с использованием алгоритмов обработки. Это встроенный оператор и он может содержать субпроцесс, состоящий из множества операторов. В process01 этот вложенный оператор содержит другие операторы внутри. При двойном клике по нему видны субпроцессы, которые состоят из 6 операторов и связаны серийно. Их задача – преобразовать данные так, чтобы их было легко обрабатывать такими техниками data mining, как ассоциации и кластеризация. В программе справа указаны параметры этих субоператоров [5]: 1) разбивка на лексемы (под лексемой здесь понимается ассоциативная группа, состоящая из нескольких слов); 2) языковая разбивка; 3) фильтр стоп-слов; 4) фильтр знаков по длине; 5) стемминг; 6) преобразование случаев.

Оператор для process01 *text nominal* преобразует текстовые данные в номинальные или категориальные. Затем оператор *numerical to binomical* преобразовывает их в биноминальную форму. Следующий этап – сама кластеризация массива текстовых документов. Сначала нужно импортировать файл proccess02, в котором операторы похожи на process01. Все начинается с процесса обработки документов, только метод создания вектора другой. Этот метод приводит к вычислению относительных частот для каждого из терминов в каждом из документов в наборе данных. В process02 метод сокращения такой же, как в файле process01, однако значение наименьшего процента параметра отличается – здесь он 20%.

Первый метод является вычислительно более затратным, чем второй, а это значит, что выявление ассоциаций занимает куда больше времени работы, чем кластеризация. Кластерный анализ текста идет в следующем порядке: оператор берет весь набор данных и преобразует его в новый, выбрав столбцы только с числовыми значениями. Это преобразование нужно, чтобы кластеризация шла численным методом. Первый результат – это список слов, возникающий во время обработки документа, второй – пример набора, который возникает от оператора «fp growth», конечный результат – работа оператора установления ассоциаций.

В следующем этапе text mining мы проанализировали, какие слова во скольких документах обнаруживаются. Затем мы работали со вкладкой «список слов», столбец «входящие документы», где могли менять порядок убывания [8]. Прежде всего, мы указали набор текстовых данных, который обрабатывался в process01. Мы нажали на process documents, указали путь к нашим текстам и получили результаты: заголовки 5235 записей. Второй результат – набор примеров, который содержит данные состава термина каждого доку-

мента в базе данных. Третий результат – это центроиды кластерной модели.

В нашем исследовании мы также использовали язык программирования R для обработки текстовой информации. Одним из преимуществ языка R является наличие для него многочисленных расширений или пакетов (полутора тысяч доступных на CRAN пакетов) [8], которые скачиваются бесплатно напрямую из R командой install. packages(). При установке языка R на компьютер несколько базовых пакетов уже имеются в наличии: необходимый пакет base, пакет grDevices, который управляет выводом графиков, пакет cluster для специального кластерного анализа.

Анализируемый нами язык программирования R представлен как статистическая система анализа, которую создали Росс Ихак и Роберт Гентлеман [9]. Наш выбор обусловлен тем, что язык программирования R – это одновременно и язык программирования, и программное обеспечение [10]. Его привлекательные свойства заключаются в эффективной обработке данных и в простых средствах для сохранения полученных результатов исследования. Одновременно он выступает набором операторов для обработки текстовых массивов, матриц, а также иных сложных конструкций. Так, язык программирования R [10] дает возможность пользователю применять операторы циклов для последовательного анализа нескольких наборов данных, где главная особенность R – это его гибкость.

Для проведения текстового анализа данных необходимо было собрать и упаковать нужную информацию и произвести её предварительную подготовку к работе. Важно было добиться возможности прочитать при помощи языка R приготовленные в другой программе данные. В нашей работе мы определяем текстовые данные как данные, которые можно прочитать и изменить с помощью текстового редактора (Emacs/Vi и т.д.). Мы использовали пробельные символы (пробел, табуляция и т. п.), запятые или точки с запятой в качестве разделителей текстовых данных.

Данный прием был применен к интеллектуальному анализу данных для вычисления общих характеристик выборки (центр и разброс), под которой понимается набор значений, полученных в результате ряда произошедших измерений. Как центр чаще выступают среднее и медиана, а как разброс, – стандартное отклонение и квартили. Отличие среднего от медианы изначально заключается в том, что среднее хорошо функционирует в случае, когда распределение данных близко к нормальному. Тем самым медиана не так сильно зависит от характеристики распределения. В этом отношении она более устойчива (робастна).

Заключение. Основным результатом исследования является применение технологии text mining для анализа текстов на предмет поиска в них заданной терминологии определенных предметных областей. Проведен анализ функциональных возможностей библиотек языка R, отработаны методики применения программного пакета Rapid Miner для глубинного анализа текста. Оптимальной формой анализа предметных онтологий можно признать так называемое концептуальное индексирование, имеющее место, когда идет индексация текста по понятиям, а не по словам, обсуждаемых в конкретном тексте. Также были обозначены оптимальные подходы к анализу предметных онтологий: формальный и лингвистический. Первый основан на логике предикатов первого порядка, второй базируется на усвоении естественного языка, положениях семантики, правилах построения онтологий на больших текстовых массивах (корпусах текстов).

Среди выявленных проблем можно указать избыточность использования синонимов, выражающих одни и те же понятия, многозначность слов, которая вызывает двусмысленность понимания, и омонимичность терминов.

В прикладном плане применение интеллектуального анализа текстовой информации в специализированных областях в системе ЭП заключается в том, что органы власти должны позаботиться, чтобы были сформированы терминологические онтологии по всем отраслям и, соответственно, была возможность сопоставления специализированных онтологий. ЭП должно быть централизованным, так как активными пользователями данных онтологий выступают экспертные системы разных сфер ИКТ, профессиональные переводчики, сопровождающие международные конференц-переводы, государственные служащие всех ветвей власти, в связи с чем вопрос стыковки терминов и сопоставления терминосистем является чрезвычайно актуальным.

Результаты проведенной работы носят теоретический и прикладной характер и изложены в публикациях автора. Прогнозные предположения о развитии объекта исследования – это стыковка различных онтологий на основе технологий искусственного интеллекта для изучения разрозненности русскоязычных и англоязычных терминов, которые, в итоге, при сравнении с такими же построениями должны быть идентичны.

Список литературы

1. Добров, Б. В. Онтологии и тезаурусы: модели, инструменты, приложения / Б. В. Добров, В. В. Иванов, Н. В. Лукашевич, В. Д. Соловьев. – М.: Бином. Лаборатория знаний, 2009. – 173 с.
2. Макаревич, Т. И. English for ICT Students = Английский язык для изучающих информационно-коммуникационные технологии: пособие: в 2-х ч. / Т. И. Макаревич, И. И. Макаревич. – Минск: Акад. упр. при Президенте Респ. Беларусь, 2012. – 382 с.
3. Piatetsky-Shapiro, G. Knowledge Discovery in Databases / G. Piatetsky-Shapiro, W. Frawley. – New York: AAAI/MIT Press, 1991. – 168 p.
4. Ландэ, Д. В. Подход к созданию терминологических онтологий / Д. В. Ландэ, А. А. Снарский // Онтология проектирования. – 2014. – № 2(12). – С. 83–91.
5. Hofmann, M. RapidMiner: Data Mining Use Cases and Business Analytics Applications / M. Hofmann, R. Klinkenberg. – New York: Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, 2013. – 525 p.
6. Тезаурус информационно-поисковый одноязычный: Правила разработки: структура, состав и форма представления // ГОСТ 7.25.-2001. Система стандартов по информации, библиотечному и издательскому делу. – Минск: Межгосударственный совет по стандартизации, метрологии и сертификации, 2001.
7. Guarino, N. Ontologies and Knowledge Bases: Towards a Terminological Clarification / N. Guarino, P. Giaretta // Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing. – Amsterdam: IOS Press, 1995. – P. 25–32.
8. Sowa, J. Knowledge Representation: Logical, Philosophical, and Computational Foundations / J. Sowa // Brooks Cole Publishing Co., Pacific Grove, CA. 2000. – V. 45(2). – P. 61–65.
9. Ihaka, R. R: A Language for Data Analysis and Graphics / R. Ihaka, R. Gentleman // Journal of Computational and Graphical Statistics. – 1996. – Vol. 5. – № 3. – P. 299–314.
10. Matloff, N. The Art of R Programming. A Tour of Statistical Software Design / N. Matloff. – San Francisco: No Starch Press. – 2011. – 316 p.

References

1. Dobrov B. V. Ontologii i tezaurusy: modeli, instrumenty, prilozheniya [Ontologies and Thesauruses: Models, Instruments, Applications]. Moscow, Binom. Laboratoriya znanij, 2009. 173 p. (in Russian).
2. Makarevich T. I., Makarevich I. I. English for ICT Students: textbook: in 2 parts. Minsk: Academy of Public Administration under the aegis of the President of the Republic of Belarus, 2012. 382 p.
3. Piatetsky-Shapiro G., Frawley W. Knowledge Discovery in Databases. NY: AAAI/MIT Press, 1991. 168 p.
4. Lande D. V. An Approach to Creating Terminological Ontologies. Ontologiya proektirovaniya [Ontology Project Development], 2014, № 2(12), pp. 83–91 (in Russian).
5. Hofmann M., Klinkenberg R. RapidMiner: Data Mining Use Cases and Business Analytics Applications. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. 1st ed, 2013. 525 p.
6. Tezaurus informatsionno-poiskovyi odnoyazychnyi: Pravila razrabotki: sruktura, sostav I forma predstavleniya // GOST 7.25.-2001. Sistema standartov po informatsii, bibliotechomu i izdatelskomu delu [Thesaurus Information-Retrieval Monolingual: Rules for Development: Structure, Content and Display Format // GOST 7.25.-2001. System of Standards in Information, Bibliography and Publishing]. Minsk, CIS Council for Standardization, Metrology and Certification Intergovernmental, 2001 (in Russian).
7. Guarino N., Giaretta P. Ontologies and Knowledge Bases: Towards a Terminological Clarification. Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing. Amsterdam, IOS Press, 1995, pp. 57–70.
8. Sowa J. Knowledge Representation: Logical, Philosophical, and Computational Foundations. Brooks Cole Publishing Co., Pacific Grove, CA, 2000, V. 45(2), pp. 61 – 65.
9. Ihaka R., Gentleman R. A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics, 1996, Vol. 5, No 3, pp. 299–314.
10. Matloff N. The Art of R Programming. A Tour of Statistical Software Design. San Francisco, No Starch Press, 2011. 316 p.

Received: 23.04.2019 *Поступила: 23.04.2019*