УДК 51-7
ББК 22.18

МЕТОДЫ АНАЛИЗА ТЕРМИНОЛОГИЧЕСКОЙ СТРУКТУРЫ ПРЕДМЕТНОЙ ОБЛАСТИ (НА ПРИМЕРЕ МЕТОДОЛОГИИ)

Губанов Д. А.[1], Макаренко А. В.[2], Новиков Д. А.[3]
(*ФГБУН Институт проблем управления РАН, Москва*)

Предлагается автоматизированный экспертный подход к синтезу и анализу терминологической структуры предметной области. Одним из ключевых отличий разработанной методики является наличие в её составе операций формального анализа, направленного на получение количественных оценок, характеризующих терминологическую структуру изучаемой предметной области. Базовые возможности разработанного подхода продемонстрированы на примере общей методологии.

Ключевые слова: терминологическая структура, теория графов, методология

1. Введение

В соответствии с определением, приведенным в [20], *теория* – форма достоверного научного знания о некоторой совокупности объектов, представляющая собой систему взаимосвязанных утверждений и доказательств и содержащая методы объяснения и предсказания явлений и процессов некоторой *предметной области*, то есть всех явлений и процессов, описываемых данной теорией.

Изложение научных результатов, полученных в той или иной предметной области, ведется, как правило, на соответст-

[1] *Дмитрий Алексеевич Губанов, к.т.н., с.н.с.* (*dmitry.a.g@gmail.com*)
[2] *Андрей Викторович Макаренко, к.т.н., с.н.с.* (*avm.science@mail.ru*)
[3] *Дмитрий Александрович Новиков, член-корр. РАН, зам. директора* (*novikov@ipu.ru*)

вующем *профессиональном языке* [19], использующем, помимо общеупотребительной, свою специальную *терминологию*.

То есть каждой предметной области можно условно поставить в соответствие (о том, как это сделать, речь пойдет ниже) *множество терминов*, характеризующих эту предметную область и используемых в ней. Совокупность этих терминов и связей между ними (о видах связей также речь пойдет ниже) будем называть *терминологической структурой* предметной области. Отметим, во-первых, что в отличие от теории *терминологических систем* (см., например, [1, 2, 10, 11, 27, 31]), нас интересуют не общие качественные закономерности формирования и развития терминологии определенной отрасли знания, а количественные свойства ее текущего «среза», например такие, как в работе [25], для предметной области «Сетецентризм». Во-вторых, в отличие от исследуемых в рамках искусственного интеллекта методов и средств построения *онтологий* предметных областей, делающих акцент на отношения между терминами, мы рассматриваем построение терминологической структуры с точки зрения эксперта некоторой предметной области.

Ключевыми являются два вопроса. Первый вопрос – насколько «хорошо» то или иное множество терминов характеризует заданную предметную область и как построить наиболее адекватное множество терминов. Второй вопрос – можно ли, например, эти термины «упорядочить» по общности, важности, встречаемости в различных сочетаниях и т.д., и как это сделать.

Первый вопрос в статье почти не рассматривается (за исключением примера в разделе 4.1). Что касается второго вопроса, то существуют две крайности в ответе на него: в формулировке «как» – *экспертный подход* (когда человек или группа людей, являющихся экспертами в данной предметной области, выделяют множество терминов и анализируют связи между ними) и *автоматический подход* (когда перечисленные операции осуществляются компьютером по заданным алгоритмам). Преимущества и недостатки и экспертного, и автоматического подходов очевидны. Оптимум лежит, как всегда, посередине – в их балансе, т.е. достигается при применении *автоматизированного экспертного подхода* (компьютерные методы осуществля-

ют поддержку деятельности экспертов), которому мы и будем следовать в настоящем исследовании.

Предположим, что имеется (на сегодня, как правило, созданный экспертами) «*тезаурус*» – словарь, собрание сведений, корпус или свод, полномерно охватывающие понятия, определения и термины специальной области знаний или сферы деятельности, что должно способствовать правильной лексической, корпоративной коммуникации (проще говоря – пониманию в общении и взаимодействии лиц, связанных одной дисциплиной или профессией); в современной лингвистике – особая разновидность словарей общей или специальной лексики, в которых указаны семантические отношения (синонимы, антонимы, паронимы, гипонимы, гиперонимы и т.п.) между лексическими единицами [32]. Общие подходы к разработке и анализу тезаурусов кратко описаны во втором разделе настоящей работы.

Проще говоря, тезаурус описывает определенную область знаний путем перечисления всех ее основных понятий и семантических отношений между ними. В своей простейшей форме тезаурус состоит из списка важных терминов и семантических отношений между ними (ассоциативных и иерархических).

Введём в рассмотрение следующий формат тезауруса терминологической структуры предметной области (см. рис. 1а). Исходно он состоит из набора статей для определяемых терминов.

Необходимо отметить, что описание термина является более широким по содержанию, нежели его определение. Более того, принято, что определение есть часть описания, т.е. определение полностью входит в описание, а не описание является дополнительным к определению (см. рис. 1а). Поэтому определяющие термины всегда содержатся среди терминов описывающих (см. рис. 1в). При этом определяющие термины могут как входить в состав определяемых, так и выходить за границы изучаемой предметной области (пример – базовые философские категории).

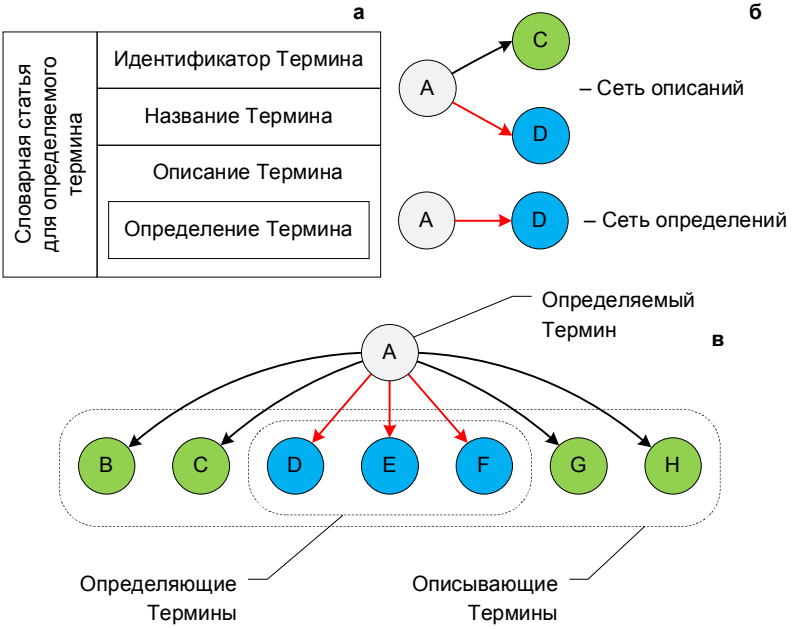

Рис. 1. Описание формата и элементов тезауруса терминологической структуры предметной области

Таким образом, между терминами вводится единственное отношение: «ссылаться» друг на друга. Это позволяет построить две сети (см. рис. 1б):

– сеть описаний – ориентированный граф, вершинами которого являются описываемые термины, а дугами – ссылки на другие термины, используемые в словарном описании первых. При этом связь направлена от описываемого термина к описывающему;

– сеть определений – является подграфом сети описаний (определение термина является частью посвященной ему «статьи» словаря).

Сеть описаний и сеть определений отражают *систему основных понятий* соответствующей предметной области. *Системность* и критерии выделения *основных понятий* требуют

детализации – ниже рассматриваются общие требования к ним (раздел 3) и приводятся примеры анализа терминологической структуры такой предметной области, как общая методология [19] (раздел 4).

Далее в разделе 2 приводится обзор подходов к разработке и анализу тезаурусов. Читатель, хорошо ориентирующийся в данной области, может перейти сразу к разделу 3.

2. Подходы к разработке и анализу тезаурусов

2.1. ОБЩИЕ ПОЛОЖЕНИЯ

Исследователями уже достаточно давно разрабатываются методики по построению и использованию тезаурусов [12, 36, 37]. На настоящий момент выработан ряд стандартов, регламентирующих формат представления тезаурусов и предоставляющих рекомендации по их созданию [см. напр. 42]. Много усилий предпринимается по созданию методов автоматического построения тезауруса, основанных на уже ставшей традиционной обработке текстов [39, 43, 45], на извлечении веб-структур (одном из направлений *Web Mining*) [38], а также на анализе онлайновых энциклопедий [47].

В России вопросами разработки тезаурусов занимаются различные научные группы, в их числе:

– Добров Б.В., Лукашевич Н.В. и др. Научно-исследовательский вычислительный центр МГУ им. М.В.Ломоносова; АНО Центр информационных исследований [3, 12];

– Филиппович Ю.Н. и др., МГТУ им. Н.Э.Баумана;

– Боровикова О.И., Загорулько Ю.А., Кононенко И.С. и др. Институт систем информатики им. А.П. Ершова СО РАН, г. Новосибирск;

– Соловьев С.Ю., Мальковский М.Г. и др., МГУ, НИУ ВШЭ [13, 14, 30];

– Рубашкин В.Ш. и др., Санкт-Петербургский государственный университет [28, 29].

2.2. СТАНДАРТЫ

На настоящий момент существует ряд стандартов, регламентирующих формат представления тезауруса, например ISO 2788 для одноязычных тезаурусов и ISO 5964 для многоязычных (соответствующие аналоги ГОСТ 7.25-2001 «Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления» и ГОСТ 7.24-90 «Тезаурус информационно-поисковый многоязычный. Состав, структура и основные требования к построению») [15]. В 2011 году стандарты ISO 2788 и ISO 5964 заменил стандарт ISO 25964-1 «Тезаурусы и совместимость с другими словарями – часть 1. Тезаурусы для информационного поиска» [42]. Разработано семейство формальных языков *SKOS* для представления тезаурусов, соответствующих парадигме *SemanticWeb* [46]. Здесь необходимо отметить, что сама по себе парадигма *SemanticWeb* в рамках целой сети Интернет подвергается активной критике, но для задач создания ограниченных словарных систем на сегодня является вполне адекватной технологией.

2.3. ОБЩАЯ МЕТОДИКА СОСТАВЛЕНИЯ И АНАЛИЗА ТЕЗАУРУСОВ

Типовая методика по разработке и анализу тезауруса состоит из следующих шагов [36, 40, 41, 44].

(Выбор терминов)

Шаг 1. Выбор источников.

Для начала необходимо определиться со множеством источников, по которым будет осуществлен сбор терминов. Источники можно разбить на две группы: *подготовленные источники* (*prearranged sources*, например классификационные схемы, тезаурусы, научные труды по терминологии области науки, энциклопедии, словари, глоссарии, оглавления учебников и учебных пособий, предметные указатели журналов, предметные указатели других публикаций в рассматриваемой области и т.п.) и *неподготовленные источники* (открытые, *open-ended sources*, например перечни поисковых запросов пользователей и профили интересов пользователей, выборки документов рассматриваемой области, перечни названий документов рассматриваемой области, обзоры документов, обсуждения со специалистами в

данной области, обзор проектов, деятельности в предметной области и т.п.). При выборе источников следует учитывать, что подготовленные источники требуют меньше усилий при сборе материала и возможно уже указывают на некоторые отношения между терминами, между терминами и понятиями, в то время как неподготовленные источники могут отражать актуальную терминологию и обеспечить более полный охват области. При выборе источников следует руководствоваться их актуальностью, полнотой и авторитетностью.

Шаг 2. Назначение кодов источникам.

Каждому выбранному источнику присваивается идентификатор для отслеживания его использования при разработке тезауруса (может потребоваться при принятии решения о предпочтительности тех или иных терминов, а также при возникновении вопроса о том, откуда взят термин).

Шаг 3. Выбор терминов.

В случае с подготовленными источниками термины могут быть прямо перенесены в базу данных терминов, при этом следует принять решение о включении источника целиком или только его части. В случае с неподготовленными источниками необходимо экспертно проанализировать источник для выбора терминов, которые могут быть полезны непосредственно или в качестве ссылок на другие термины. Также можно использовать программное обеспечение для автоматизированной «добычи» ключевых слов и фраз при создании списков терминов-кандидатов и последующем выборе терминов из списка.

Шаг 4. Внесение терминов в базу данных с сопутствующей информацией.

(Объединение терминов и разработка классов концептов)

Шаг 5. Сортировка терминов базы данных в алфавитном порядке.

Шаг 6. Объединение одинаковых терминов.

Объединение (слияние) информации для одинаковых терминов (при этом может потребоваться информация из дополнительных источников).

Шаг 7. Объединение терминов в один класс концептов.

Слияние синонимов или терминов в один класс концептов.

(Определение предметных областей и подобластей)

В самом начале анализа терминологии предметной области маловероятно, что аналитик обладает достаточным уровнем знаний для детального структурирования предметной области. Поэтому, как правило, применяется подход «сверху-вниз» [44], когда сначала рассматриваются более общие предметные области, затем их подобласти и так далее.

Шаг 8. Определение общих предметных областей.

На данном шаге необходимо определить общие предметные области и распределить в них термины.

Шаг 9. Определение подобластей в одной предметной области.

На данном шаге необходимо определить подобласти каждой общей области и распределить термины в эти подобласти.

Шаг 10. Разработка детальной структуры предметной области.

На данном шаге необходимо выбрать предпочтительные термины (среди терминов, относящихся к одному понятию, выделяют наиболее подходящий термин, который наиболее хорошо характеризует или обозначает данное понятие) и выполнить слияние (объединение) информации для терминов в одном классе концептов.

(Разработка классификационной структуры)

Шаг 11. Улучшение структуры классов.

На данном шаге необходимо разработать предварительный вариант систематического указателя терминов[4] (или внести исправления в существующий вариант систематического указателя) и обновить рабочую базу данных.

Шаг 12. Рецензирование систематического указателя.

Затем необходимо проверить то, как обстоит дело «на практике»: необходимо передать текущий вариант систематического указателя пользователям и экспертам предметной области.

(Критический анализ)

[4] *Систематический указатель* (*информационно-поискового*) *тезауруса – вспомогательная часть информационно-поискового тезауруса, в которой перечень лексических единиц* (*терминов*) *построен согласно с принятой классификацией понятий соответствующей отрасли знания.*

Шаг 13. Обсуждение систематического указателя с экспертами и пользователями.

Необходимо обсудить систематический указатель с пользователями и экспертами, в случае большого количества замечаний перейти на шаг 11. Затем выработать полную версию тезауруса, проверить перекрестные ссылки, при необходимости вставить перекрестные ссылки.

(Тестирование)

Шаг 14. Тестирование тезауруса.

Для тестовой выборки документов необходимо назначить дескрипторы (т.е. предпочтительные термины), определить наличие соответствующих терминов в тезаурусе.

(Ревизия тезауруса)

Шаг 15. Ревизия тезауруса.

Необходимо установить регулярный график обзоров и пересмотров тезауруса, обеспечить сбор «жалоб и предложений» со стороны пользователей.

3. Подход к разработке и анализу терминологической структуры предметной области

3.1. МЕТОДИКА СИНТЕЗА И АНАЛИЗА ТЕРМИНОЛОГИЧЕСКОЙ СТРУКТУРЫ ПРЕДМЕТНОЙ ОБЛАСТИ

Методика составления и анализа тезаурусов, представленная в разделе 2.3, была переработана авторами с целью формирования автоматизированного экспертного подхода к синтезу и анализу терминологической структуры предметной области. В итоге получена достаточно общая методика, схема которой изображена на рис. 2.

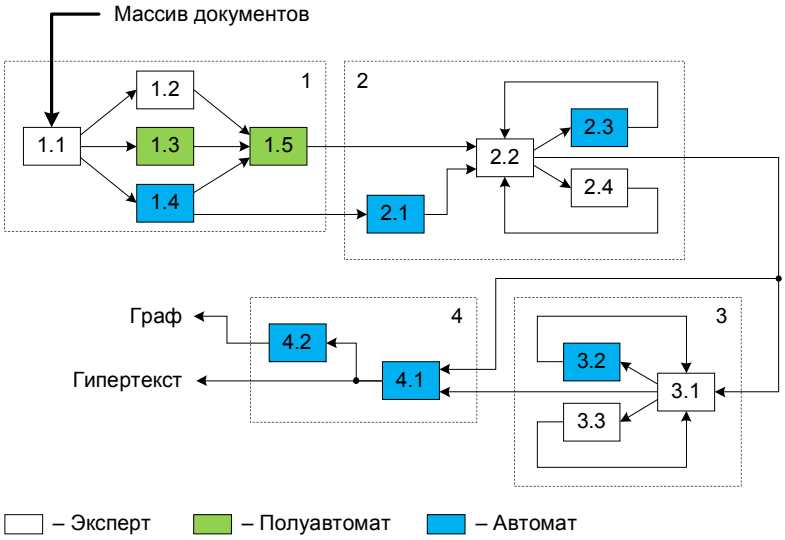

Рис. 2. Методика разработки и анализа терминологической структуры предметной области

Как видно из рис. 2, методика состоит из четырёх основных блоков:

1. Выделение совокупности основных терминов предметной области.
2. Синтез и анализ сети описаний.
3. Синтез и анализ сети определений.
4. Представление результатов анализа.

По степени автоматизации операции, входящие в состав описываемой методики, делятся на три вида: выполняются полностью вручную (Эксперт); труд эксперта частично автоматизирован (Полуавтомат); полностью автоматически (Автомат).

Блок 1 включает в себя пять операций:

1.1. Подбор исходного корпуса текстов, отражающих современное состояние данной теории.

1.2. Выделение множества необщеупотребительных терминов, принадлежащих предметной области.

1.3. Анализ структуры библиографических ссылок.

1.4. Выделение множества необщеупотребительных терминов, принадлежащих выбранному корпусу текстов.

1.5. Анализ и построение совокупности основных терминов предметной области, формирование их определений.

Блок 2 включает в себя четыре операции:

2.1. Построение лексической сети.

2.2. Создание «словаря» терминов (включая взаимные ссылки).

2.3. Формальный анализ сети описаний.

2.4. Содержательный анализ сети описаний.

Блок 3 состоит из трёх операций:

3.1. Построение и верификация сети определений.

3.2. Формальный анализ сети определений.

3.3. Содержательный анализ сети определений.

Блок 4 состоит из двух операций:

4.1. Создание гипертекстовой версии «словаря».

4.2. Визуализация сети описаний и сети определений.

В дальнейшем пользователь (специалист или человек, осваивающий предметную область, пользуется конечными результатами – гипертекстовой версией «словаря» и соответствующими средствами визуализации).

Необходимо отметить, что разработанная методика синтеза и анализа терминологической структуры предметной области изложена на уровне общей схемы. Каждая операция предполагает использование различных инструментов, методов, средств и алгоритмов из нескольких областей науки, в том числе: компьютерная лингвистика [37], математическая статистика [6], теория графов [34], программирование [8].

Анализ существующих подходов (см., например, [4]) и собственный опыт авторов статьи показывает, что экспертная деятельность по созданию собственно словаря (операция 2.2.) довольно трудоемка и пока очень слабо поддаётся автоматизации. В этом плане сделаем ряд общих замечаний:

1. Обеспечение корректности сети определений (например, ацикличность графа определений является важнейшим показателем непротиворечивости системы определений - см. также другие требования к системе определений и правила определения понятий в [1, 7]) может потребовать неоднократного циклического возврата к предыдущим этапам.

2. Целесообразно выделение как минимум трех «уровней» терминов: базовые философские категории, базовые категории предметной области, частные понятия предметной области.

3. Удобным приемом анализа является выделение «окрестности» термина – множества терминов, длина пути от/до которых в сети описаний (или в сети определений – в зависимости от целей анализа) не превышает заданной.

Наличие в составе методики блока 4, включающего средства визуализации, позволяет не только наглядно представить результаты синтеза, но и значительно упрощает анализ и изучение терминологической структуры предметной области. Другими словами, визуализация является не столько инструментом создания и анализа последней, сколько подспорьем для неспециалиста, желающего быстро и самостоятельно составить для себя целостную картину или ищущего ответ на некоторый конкретный частный вопрос.

3.2. ФОРМАЛЬНЫЙ АНАЛИЗ ТЕРМИНОЛОГИЧЕСКОЙ СТРУКТУРЫ ПРЕДМЕТНОЙ ОБЛАСТИ

Одним из ключевых отличий данной методики является наличие в её составе двух операций (см. рис. 2): 2.3 – «Формальный анализ сети описаний» и 3.2 – «Формальный анализ сети определений». Формальный анализ направлен на получение количественных оценок, характеризующих терминологическую структуру изучаемой предметной области.

Исходя из того, что сеть описаний и сеть определений представлены в виде графов (вершиной является термин, а дугой – связь между терминами, см. рис. 1б и 1в), основным инструментом исследования структуры этих сетей являются различные топологические и метрически характеристики графов, а также величины, определённые над этими характеристиками.

Необходимо отметить, что каждая из сетей (сеть описаний и сеть определений) представляются связными ориентированными графами без петель и без кратных дуг. Кроме того, граф определений ацикличен, т.е. не содержит циклов и между парами вершин имеется только по одному пути.

Таким образом, для каждой из сетей рассчитываются следующие *первичные показатели*:

In_i – число входящих дуг;

Out_i – число исходящих дуг;

$PageRank_i$ – вероятность попадания в данную вершину при случайном блуждании по дугам сети;

$ClosIn_i$ – близость вершин сети к данной вершине;

$ClosOut_i$ – близость данной вершины ко всем остальным вершинам сети;

$Betw_i$ – характеризует расположение вершины на кратчайших путях сети (здесь и далее i – это номер вершины графа (номер термина)).

На основе первичных показателей для каждого термина определяются две производные характеристики:

– «*значимость термина*» – количественно характеризует его эффективную «используемость» в описании других терминов.

– «*сложность термина*» – количественно (условно) показывает, насколько другие термины используются в его описании.

Значимость i-го термина вычисляется по формуле ($0 \le S_i \le 1$):

$$S_i = \frac{\max R - R_i}{\max R - \min R},$$

где $R_i = \max\{\text{Rank } In_i, \text{Rank } ClosIn_i, \text{Rank } PageRank_i\}$. Здесь и далее Rank – функция вычисления ранга величины. Значимость i-го термина максимальна при $S_i = 1$.

В свою очередь, сложность i-го термина вычисляется по формуле ($0 \le C_i \le 1$):

$$C_i = \frac{\max r - r_i}{\max r - \min r},$$

где $r_i = \max\{\text{Rank } Out_i, \text{Rank } ClosOut_i, \text{Rank } Betw_i\}$. Сложность i-го термина максимальна при $C_i = 1$.

Наличие первичных метрических характеристик сети описаний и сети определений позволяет исследователю проявлять «полет фантазии» и конструировать разнообразные и, быть может, более информативные, чем «сложность» и «значимость», вторичные показатели.

Как уже было отмечено выше, граф определений ацикличен и при этом является направленным, следовательно, содержит вершины (термины) без исходящих связей (не определяемые через термины предметной области). Таким образом, все термины можно разбить на упорядоченные классы, поставив каждому из них в соответствие максимальный и минимальный «уровень» – длину соответственно максимального или минимального пути до класса «неопределяемых» терминов (этому классу условно припишем первый уровень). Пара этих длин является ещё одной информативной характеристикой, описывающей структуру сети определений.

Далее, к полученным показателям и характеристикам возможно применить широкий арсенал математических средств, направленных на изучение взаимосвязи между этими показателями и характеристиками.

3. Пример: терминологическая структура методологии

Проиллюстрируем результаты применения описанного в предыдущем разделе общего подхода к синтезу и анализу терминологической структуры предметной области, на примере методологии.

4.1. ВЫДЕЛЕНИЕ СОВОКУПНОСТИ ОСНОВНЫХ ТЕРМИНОВ ПРЕДМЕТНОЙ ОБЛАСТИ

Операция 1.1. Подбор исходного корпуса текстов. В качестве исходного корпуса текста использовалась объёмная монография [19], в которой приведено наиболее полное, на сегодняшний день, изложение содержания общей методологии.

Операция 1.2. Выделение множества необщеупотребительных терминов. Если проанализировать совокупность используемых в упомянутой монографии необщеупотребительных терминов (т.е. всего множества слов текста исключая служебные и общеупотребительные слова русского языка), то, во-первых, 200 первых по числу упоминаний терминов дают около 90% суммарного числа упоминаний всех терминов. Во-вторых, оказывается, что словарь [21] охватывает 80–90% наиболее

часто встречающихся из них (другими словами, совокупность терминов словаря достаточно полно покрывает терминологию такой предметной области, как общая методология) – верхняя прямая на рис. 3 (обозначена $\hat{I}$) является биссектрисой, нижняя «прямая» (I_j) показывает число терминов словаря, входящих в упорядоченное по убыванию встречаемости в [19] (Q_j) множество терминов методологии. Другими словами, различие между этими прямыми показывает, насколько словарь не соответствует исходному корпусу текстов.

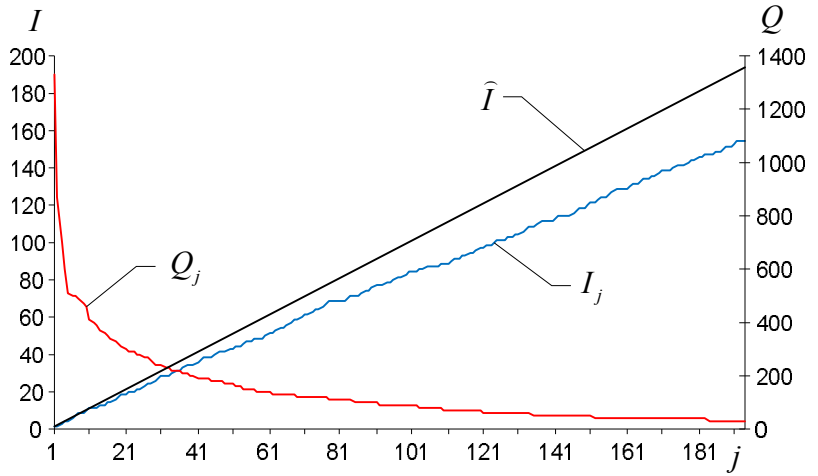

Рис. 3. Распределение терминов [*19*] *по числу упоминаний и соответствие со словарем* [*21*]

Кривая Q_j, $j = 1, \ldots, 194$, с достаточно высокой степенью точности аппроксимируется функцией (средняя относительная ошибка ~ 4%):

$$\tilde{Q}[j] = \frac{256,605 - 249,079\sqrt{j}}{0,111117 - 0,105957\, j} - 126,143 .$$

Необходимо отметить, что при значении $j = \hat{j} = 310$, функция $\tilde{Q}[j]$ впервые становится отрицательной. Таким образом, количество терминов на уровне $\hat{j}$ возможно определить, как

некое граничное, отражающее исчерпание списка основных (!) терминов предметной области (магическое число «~300» как типовая емкость тезауруса или словаря предметной области будет еще неоднократно упоминаться нами ниже).

Операция 1.3. Анализ структуры библиографических ссылок. Совокупность цитирований в [21] типична для любой предметной области (см. описание общих закономерностей в [35]). Сводная информация приведена на рис. 4.

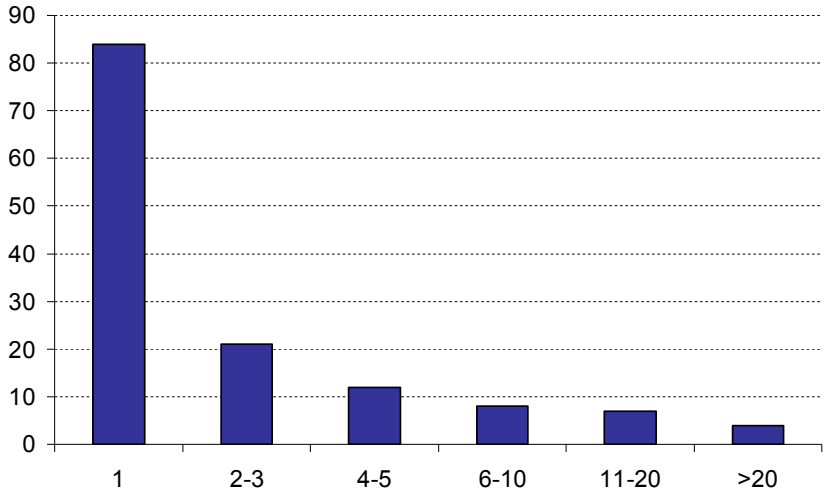

Рис. 4. Распределение литературных источников [*21*] *по числу упоминаний*

Большинство терминов заимствовано из работ по методологии [19, 20] (39 и 69 упоминаний соответственно), системному анализу [26] (29 упоминаний), а также из классических словарей по философии [16, 33] (соответственно 17 и 36 упоминаний), логике [7] (17 упоминаний), психологии [9] (11 упоминаний) и других профессиональных словарей и энциклопедий. Большинство работ (76) упоминаются только один раз, 2–4 раза упоминаются 28 работ, 9 работ упоминаются от 5 до 10 раз.

Итак, анализ структуры библиографических ссылок позволяет не только выделить «основные» работы, но и оценить «количественно» связь исследуемой предметной области с другими предметными областями.

Операция 1.4. Анализ совокупности основных терминов предметной области. Характерным (хотя и субъективно зависящим от эксперта – составителя глоссария) числом относительно часто встречающихся в предметной области необщеупотребительных слов является 250–500. В этот диапазон укладываются как результаты автоматического анализа [19] (300 терминов дают 95% суммарного числа упоминаний всех терминов), так и экспертно составленный словарь [21] – 297 терминов, а также экспертно составленные словарь по педагогике [18] (около 300 терминов), включенный в [22] глоссарий по теории управления организационными системами (около 300 терминов – см. также http://www.mtas.ru/Glossary.htm), глоссарий по интеллектуальным системам на сайте www.glossary-ipu.ru (490 терминов).

Наиболее часто в [19] встречаются следующие термины (перечисление в порядке убывания частоты – см. убывающую кривую, построенную по вспомогательной (правой) оси, на рис. 3, а также облако тегов по адресу http://www.mtas.ru/theory/methodology/Cloud.png):

(1) деятельность, научный, система, исследование, процесс, проект, метод, результат, знание, игра, наука, работа, обучение, проблема, художественный, теория, модель, форма, задача, организация, решение, методология, общий, развитие, образ, объект, средство, оценка, условие, принцип.

Из этих 30 терминов 27 входят в словарь [21] – см. рис. 3 (не вошли в словарь «обучение», «художественный» и «общий» – эксперты не сочли эти термины характерными для методологии). Таким образом, автоматические и экспертные методы выделения совокупности основных терминов предметной области хорошо согласуются по результатам и дополняют друг друга.

4.2. СИНТЕЗ И АНАЛИЗ СЕТИ ОПИСАНИЙ

Операция 2.1. Построение лексической сети. В данном исследовании полноценная лексическая сеть не строилась, ибо анализ проводился на уровне униграмм.

Операция 2.2. Создание «словаря» терминов. Как отмечалось выше, словник словаря [21] составлялся авторами «вручную», описания каждого из 297 терминов составлялись без привлечения «средств автоматизации». В описании каждого термина вручную выделялась часть, соответствующая определению термина и указывались ссылки на другие термины словаря, что позволило автоматически сформировать сеть описаний и сеть определений.

Операция 2.3. Формальный анализ сети описаний. Имея сеть описаний, вычислим для каждого *i*-го термина (узла сети) значения вышеперечисленных первичных показателей и производные от них характеристики – сложность и значимость, $i = 1, \ldots, 297$ (исходные данные и первичные показатели содержатся в *XLS*-файле (http://www.mtas.ru/theory/methodology/DescriptionsNetwork.xls), что позволяет заинтересованному читателю самостоятельно исследовать интересующие его свойства сети описаний и сети определений методологии).

Высокие ($> 0{,}8$) коэффициенты статистически значимой ранговой корреляции Спирмена[5] имеют место между тремя переменными – *In*, *ClosIn* и *PageRank*. Первые 30 терминов (упорядоченных по убыванию значимости) таковы (ср. с (1)):

(2) деятельность, процесс, объект, форма, система, цель, явление, метод, знания, теория, действие, результат, организация, наука, элемент, средство, проблема, свойство, способ, познание, субъект, условие, исследование, содержание, единство, предмет, принцип, операция, понятие, задача.

Коэффициент корреляции Спирмена между значимостью термина и частотой его встречаемости в исходном корпусе текстов превышает 0,75.

Первые 30 терминов (упорядоченных по убыванию сложности) таковы (ср. с (1) и (2)):

[5] *В работе принят уровень значимости* $\alpha = 10^{-4}$, *граничное значение коэффициента корреляции Спирмена равно* $\sim \pm 0{,}226136$ [6].

(3) деятельность, теория, метод, наука, моделирование, проект, системный анализ, проблема, эксперимент, прогнозирование, экспертные оценки, рефлексия, метод сценариев, методология, средство, научная деятельность, семиотика, подход, опытная работа, гипотеза, организационная культура, управление проектами, оптимизация, научные знания, метод мозгового штурма, формы организации научного знания, проектирование, норма, кибернетика, понятие.

Для сети описаний статистически значимая ранговая корреляция между сложностью и значимостью, а также между сложностью и частотой его встречаемости в исходном корпусе текстов практически отсутствует (коэффициент корреляции Спирмена < 0,2). На рис. 5 приведёно соответствие между значениями параметров «Значимость» и «Сложность» для сети описаний.

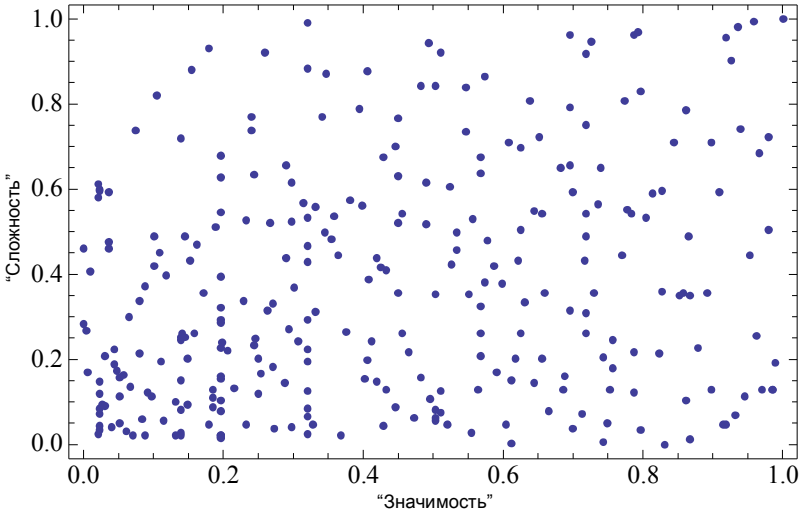

Рис. 5. Соответствие между значениями параметров «Значимость» и «Сложность» для сети описаний

Кластерный анализ сети описаний не позволил разделить термины на различающиеся содержательно интерпретируемые кластеры.

На рис. 6 для сети описаний приведены графики величин «Значимость» и «Сложность», упорядоченных по убыванию. Видно, что обе эти характеристики изменяются достаточно равномерно, т.е. на их основании трудно выделить в сети описаний то или иное характерное множество терминов.

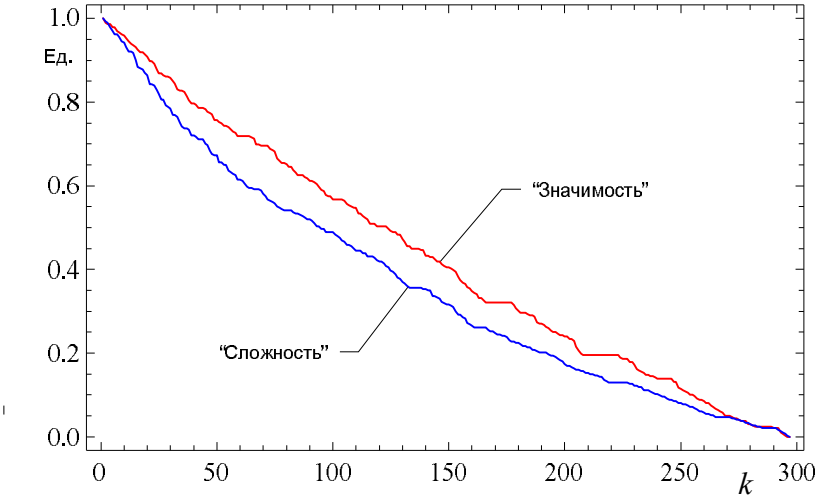

Рис. 6. Графики величин «Значимость» и «Сложность», упорядоченных по убыванию, для сети описаний

Операция 2.4. Содержательный анализ сети описаний заключается, например, в экспертном анализе упорядочений (1)–(3).

4.3. СИНТЕЗ И АНАЛИЗ СЕТИ ОПРЕДЕЛЕНИЙ

Операция 3.1. Построение и верификация сети определений. Автоматическая верификация показала, что сеть определений ациклична.

Операция 3.2. Формальный анализ сети определений. Как было указано в разделе 3.2, аналогично сети описаний, для каждого термина в сети определений можно вычислить шесть первичных показателей и два производных (исходные данные и

первичные показатели содержатся в *XLS*-файле (http://www.mtas.ru/theory/methodology/DefinitionsNetwork.xls)).

Первые 30 терминов (упорядоченных по убыванию значимости) таковы (ср. с (1)–(2)):

(4) деятельность, процесс, явление, знания, результат, действие, объект, познание, система, свойство, субъект, цель, задача, предмет, операция, элемент, поведение, условие, проблема, единство, метод, потребность, проект, фактор, альтернатива, правило, сознание, решение, управление, показатель.

Первые 30 терминов (упорядоченных по убыванию сложности) таковы (ср. с (1)–(4)):

(5) логика, теория, научные знания, наука, эстетика, дедуктивный метод, аналогия, рефлексия, таксономия, имитационное моделирование, институт, теорема, практика, планирование, метод экспертных оценок, научно-исследовательская работа, аналог, иерархия, эксперимент, доказательство, положение, опыт, предметная область, аксиома, опрос, величина, семантика, мораль, системный анализ, содержание.

Для сети определений существует небольшая (коэффициент корреляции Спирмена равен – 0,241) статистически значимая отрицательная ранговая корреляция между сложностью и значимостью.

На рис. 7 для сети определений приведёно соответствие между значениями параметров «Значимость» и «Сложность». В отличие от сети описаний (см. рис. 5), в сети определений присутствует некая «запрещённая» область (см. рис. 7). Её наличие можно условно интерпретировать как некую изотропность («от простого – к сложному») выделенного терминологического ядра методологии – в системе определений отсутствуют сложные и одновременно значимые термины.

На рис. 8 для сети определений приведены графики величин «Значимость» и «Сложность», упорядоченных по убыванию. Из рис. 8 видно, что имеется значительное число терминов, обладающих высокой одинаковой сложностью (плато А). При этом почти половина терминов имеет минимальную значимость (плато В).

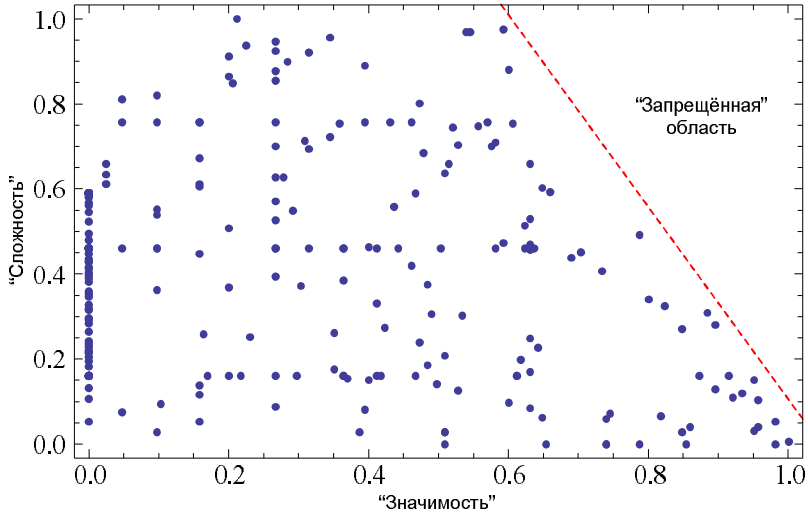

Рис. 7. Соответствие между значениями параметров «Значимость» и «Сложность» для сети определений

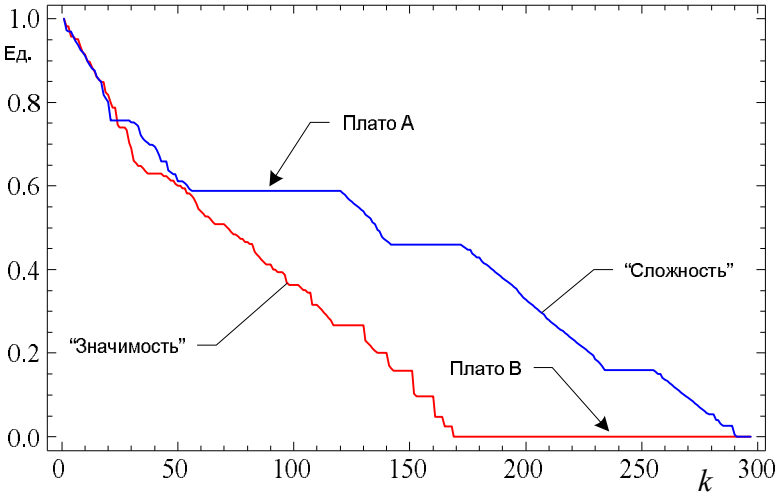

Рис. 8. Графики величин «Значимость» и «Сложность», упорядоченных по убыванию, для сети определений

Также существует статистически значимая положительная ранговая корреляция (коэффициент корреляции Спирмена равен 0,659) между значимостью термина и частотой его встречаемости в исходном корпусе текстов.

Кластерный анализ сети определений не позволил разделить термины на различающиеся содержательно интерпретируемые кластеры.

Анализ сети определений (см. раздел 3.2.) позволил выделить семь терминов без исходящих связей (т.е. не определяемых через термины предметной области):

альтернатива, оценка, потребность, сознание,
требование, элемент, явление.

Оказывается, что при классификации по длине минимального пути сеть определений можно разбить на шесть уровней, а при классификации по длине максимального пути – на сорок (!) уровней (см. рис. по адресу http://www.mtas.ru/theory/methodology/Hierarchy.png).

Для сети определений существует статистически значимая положительная ранговая корреляция (коэффициент корреляции Спирмена равен 0,731) между сложностью термина и его максимальным уровнем. Также существует статистически значимая отрицательная ранговая корреляция (коэффициент корреляции Спирмена равен –0,480) между значимостью термина и его максимальным уровнем. Кроме того, следует отметить статистически значимую ранговую корреляцию (коэффициент корреляции Спирмена равен 0,993) между близостью термина к остальным терминам (*ClosOut*) и его максимальным уровнем.

Операция 3.3. Содержательный анализ сети определений заключается, например, в экспертном анализе упорядочений (1), (4) и (5).

4.4. ПРЕДСТАВЛЕНИЕ РЕЗУЛЬТАТОВ АНАЛИЗА

Приведенные выше количественные результаты п.п. 4.1–4.3, отраженные в виде соответствующих рисунков, таблиц, графов связей и т.д., наглядно представляют свойства структуры связей между терминами (см. рисунки сетей по адресу http://www.mtas.ru/theory/methodology/Hierarchy.png и http://www.mtas.ru/theory/methodology/DescriptionsNetwork.png).

Наличие гипертекстовой версии словаря [21] (см. http://www.mtas.ru/biblio/Methodology_g.htm) позволяют пользователю анализировать взаимосвязь терминов и получать системное представление о терминологической структуре методологии.

Представленные на сайте http://ics-social.appspot.com средства визуализации (разработанные к.т.н. Д.А. Губановым) сети описаний и сети определений позволяют наглядно представлять (в изменяемом пользователем графическом масштабе при различных размерах окрестностей выбранного термина) связи между терминами.

4. Заключение

В настоящей работе предложен автоматизированный экспертный подход к синтезу и анализу терминологической структуры предметной области, который может быть использован, во-первых, специалистами по искусственному интеллекту (анализу предметных областей, извлечению и систематизации знаний в виде тезаурусов, онтологий и т.д.). Во-вторых – специалистами в конкретных предметных областях при анализе структуры и свойств последних. В-третьих, неспециалистами, стремящимися освоить новую для себя предметную область.

Результаты (раздел 4) применения методики (раздел 3) анализа терминологической структуры (системы основных понятий) методологии представляются системными и иногда даже несколько неожиданными (примеры – соотношение между сложностью и значимостью терминов, глубина иерархической структуры терминов и др.).

В целом относительно методов и результатов формального анализа сетей описаний и определений можно сделать следующее замечание. В настоящей работе реализована традиционная для статистического исследования общая схема (см., например, [6, 23, 24]): описательная статистика (для части первичных и вторичных показателей), исследование зависимостей (значимые корреляции выявлены), снижение размерности (дисперсионный анализ и метод главных компонент не дали значимых и содержательно интерпретируемых результатов), кластерный анализ.

В зависимости от того, какие вопросы ставит перед собой исследователь, можно использовать соответствующие методы из богатого их арсенала, имеющегося в современной области *Data Mining* [48].

Перспективным представляется возможный будущий анализ терминологической структуры различных предметных областей (а также «операций» их объединения), что, конечно, потребует от исследователей определенных усилий (см. выше неавтоматизируемые, экспертные компоненты методики на рис. 2), но в значительной степени облегчается наличием уже созданных (и существующих в электронном виде) семантических сетей и онтологий во многих отраслях научного знания. Кроме того, формализация терминологических структур предметных областей может оказаться востребованной при построении автоматизированных обучающих систем.

Еще раз подчеркнем гибкость предложенного подхода – он является лишь общей схемой, в рамках которой, как элементы «конструктора», могут быть использованы разнообразные методы и алгоритмы современной теоретической и компьютерной лингвистики, теории графов, прикладной статистики, искусственного интеллекта и др.

Авторы признательны за плодотворные обсуждения к.ф.-м.н. Н.В, Лукашевич, к.т.н. Л.И. Микуличу, д.т.н. О.П. Кузнецову и д.т.н. С.Ю. Соловьеву.

Литература

1. БЛОХ М.Я. *Теоретические основы грамматики*. – М.: Высшая школа, 2004. – 237 с.
2. ГОЛОВАНОВА Е.И. *Введение в когнитивное терминоведение*. – М.: Флинта, 2011. – 224 с.
3. ДОБРОВ Б.В., ИВАНОВ В.В., ЛУКАШЕВИЧ Н.В., СОЛОВЬЕВ В.Д. *Онтологии и тезаурусы: модели, инструменты, приложения*. – М.: Интернет-университет информационных технологий, 2009. – 176 с.
4. ЗАГОРУЛЬКО Ю.А., БОРОВИКОВА О.И., КОНОНЕНКО И.С., СОКОЛОВА Е.Г. *Методологические аспекты разработки электронного русско-английского тезауруса по*

компьютерной лингвистике // Информ. и её примен. – 2012. – Т.6, В.3. – С. 22–31.
5. ИВИН А.А. *Логика*. – М.: Знание, 1998. – 240 с.
6. КОБЗАРЬ А.И. *Прикладная математическая статистика*. – М.: Физматлит, 2006. – 816 с.
7. КОНДАКОВ Н.И. *Логический словарь-справочник*. – М.: Наука, 1975. – 720 с.
8. КОРМЕН Т.Х. и др. *Алгоритмы: построение и анализ*. – М. Вильямс, 2006. – 1296 с.
9. *Краткий психологический словарь* / Сост. Л.А. Карпенко. Под общ. ред. А.В. Петровского, М.Г. Ярошевского. – М.: Политиздат, 1985. – 287 с.
10. ЛЕЙЧИК В.М. *Терминоведение*. – М.: ЛКИ, 2007. – 256 с.
11. ЛОТТЕ Д.С. *Основы построения научно-технической терминологии* // Вопросы теории и методики. – М.: Изд-во академии наук СССР, 1968. – 160 с.
12. ЛУКАШЕВИЧ Н.В. *Тезаурусы в задачах информационного поиска*. – М.: МГУ, 2011. – 512 с.
13. МАЛЬКОВСКИЙ М.Г., СОЛОВЬЕВ С.Ю. *Методы формирования глоссариев в универсальном терминологическом пространстве* // Труды Международного семинара Диалог'2003 «Компьютерная лингвистика и интеллектуальные технологии». – М.: Наука, 2003. – С. 438–440.
14. МАЛЬКОВСКИЙ М.Г., СОЛОВЬЕВ С.Ю. *Терминологические сети* // Материалы 2-й международной научно-технической конференции «Открытые семантические технологии проектирования интеллектуальных систем». – Минск: БГУИР, 2012. – С. 77–82.
15. НГУЕН М.Х., АДЖИЕВ А.С. *Описание и использование тезаурусов в информационных системах, подходы и реализация.* – [Электронный ресурс] URL: http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2004/part1/NA (дата обращения: 24.05.2012).
16. *Новая философская энциклопедия*: В 4 томах / Под редакцией В.С. Стёпина. – М.: Мысль, 2000-2001. – с.
17. НОВИКОВ А.М. *Основания педагогики*. 2-е изд. – М.: Эгвес, 2011. – 208 с.

18. НОВИКОВ А.М. *Педагогика: словарь системы основных понятий*. – М.: Издательский центр ИЭТ, 2013. – 268 с.
19. НОВИКОВ А.М., НОВИКОВ Д.А. *Методология*. – М.: Синтег, 2007. – 668 с.
20. НОВИКОВ А.М., НОВИКОВ Д.А. *Методология научного исследования*. – М.: Либроком, 2010. – 280 с.
21. НОВИКОВ А.М., НОВИКОВ Д.А. *Методология: словарь системы основных понятий*. – М.: Либроком, 2013. – 208 с.
22. НОВИКОВ Д.А. *Теория управления организационными системами*. 3-е изд. испр. и дополн. – М.: Физматлит, 2012. – 604 с.
23. НОВИКОВ Д.А., НОВОЧАДОВ В.В. *Статистические методы в медико-биологическом эксперименте* (типовые случаи). – Волгоград: Издательство ВолГМУ, 2005. – 84 с.
24. ОРЛОВ А.И. *Эконометрика*. – М.: Экзамен, 2004. – 576 с.
25. ПАВЛОВСКИЙ И.С. *Исследование сетецентрического подхода к управлению на основе однородного концептуального моделирования* // Труды и пленарные доклады участников конференции УКИ'12. - М.:ИПУ РАН, 2012. - 1 электрон. опт. диск (CD-ROM). - ISBN 978-5-91450-100-3. – С. 1699–1706.
26. ПЕРЕГУДОВ Ф.И., ТАРАСЕНКО Ф.П. *Введение в системный анализ*. – М.: Высшая школа, 1989. – 367 с.
27. РЕФОРМАТСКИЙ А.А. *Терминология. Введение в языкознание*. – М.: Учпедгиз, 1959. – 536 с.
28. РУБАШКИН В.Ш. *Представление и анализ смысла в интеллектуальных информационных системах*. — М.: Наука, 1989. – 192 с.
29. РУБАШКИН В.Ш. *Семантический компонент в системах понимания текста* // Труды Десятой национальной конференции по искусственному интеллекту с международным участием (КИИ-2006). – М.: Физматлит, 2006. - Т. 2. – С. 455–463.
30. СОЛОВЬЕВ С.Ю. *Образные представления терминологической сети* // Прикладное программное обеспечение. – М.: Изд-во МИРЭА, 2008. – С. 55–69.

31. СУПЕРАНСКАЯ А.В., ПОДОЛЬСКАЯ Н.В., ВАСИЛЬЕВА Н.В. *Общая терминология. Вопросы теории.* Изд. 4-е. – М.: Изд-во ЛКИ, 2007. – 248 с.
32. Тезаурус. - URL: *http://ru.wikipedia.org/wiki/Тезаурус*.
33. *Философский энциклопедический словарь*. – М.: Сов. Энциклопедия, 1983. – 836 с.
34. ХАРАРИ Ф. *Теория графов*. – М.: Мир, 1973. – 324 с.
35. ЯБЛОНСКИЙ А.И. *Модели и методы исследования науки*. – М.: Эдиториал УРСС, 2001. – 400 с.
36. AITCHISON J., GILCHRIST A., BAWDEN D. *Thesaurus Construction And Use*: *A Practical Manual*. – Aslib, 2000. – 233 p.
37. BIRD S., KLEIN E. LOPER E. *Natural Language Processing with Python*. – O'Reilly Media, 2009. – 504 p.
38. CHEN Z., LIU S., WENYIN L., PU G., MA W.Y. *Building a Web Thesaurus from Web Link Structure* // Proc. ACM SIGIR, 2003. – P. 48–55.
39. CROUCH C.J. A *Cluster Based Approach to Thesaurus Construction* // Proc. ACM SIGIR, 1988. – P. 309–320.
40. GREFENSTETTE G. *Explorations in Automatic Thesaurus Discovery*. – Kluwer Academic Publishers, Hingham, MA, 1994.– 278 p.
41. *Handbook of Terminology Management: Basic Aspects of Terminology Management* / Sue Ellen Wright, Gerhard Budin. – John Benjamins Pub Co, 1997. – 370 p.
42. ISO 25964-1, http://www.iso.org/iso/catalogue_detail.htm ?csnumber=53657.
43. MANNING C.D., SCHUTZE H. *Foundations of Statistical Natural Language Processing*. – MIT, 1999. – 620 p.
44. SAGER J.C. *A Practical Course in Terminology Processing*. – J. Benjamins Pub. Co., 1990. – 254 p.
45. SCHUTZE H., PEDERSEN J.O. *A Cooccurrence-based Thesaurus and Two Applications to Information Retrieval* // International Journal of Information Processing and Management. – 1997. – №33. – P. 307–318.
46. SKOS Simple Knowledge Organization System. – URL: http://www.w3.org/2004/02/skos/.

47. STRUBE M., PONZETTO S. *WikiRelate! Computing Semantic Relatedness Using Wikipedia* // Proc. National Conference on Artificial Intelligence (AAAI'06), 2006. – P. 1419–1424.
48. WITTEN I.H., FRANK E., HALL M.A. *Data Mining: Practical Machine Learning Tools and Techniques.* Third Edition. – Morgan Kaufmann, 2011. – 629 p.

METHODS TO ANALYSE TERMINOLOGICAL STRUCTURE OF SUBJECT AREA

Dmitry Gubanov, Institute of Control Sciences of RAS, Moscow, Cand.Sc., senior researcher (dimagubanov@mail.ru).
Andrey Makarenko, Institute of Control Sciences of RAS, Moscow, Cand.Sc., senior researcher (avm.science@mail.ru).
Dmitry Novikov, Institute of Control Sciences of RAS, Moscow, deputy director (novikov@ipu.ru).

Abstract: *We propose an automated expert approach to analysis and synthesis of a terminological structure of a subject area. The key novelty is that the method employs formal analytical operation to obtain numerical characteristics of terminological structure of the studied subject area. Basic features of the developed approach are illustrated by the subject area of general methodology.*

Keywords: terminological structure, graph theory, methodology.

Статья представлена к публикации
членом редакционной коллегии О.П. Кузнецовым

Поступила в редакцию 28.01.2013.
Опубликована 31.05.2013.