

МЕТОДЫ АНАЛИЗА ТЕРМИНОЛОГИЧЕСКОЙ СТРУКТУРЫ ПРЕДМЕТНОЙ ОБЛАСТИ (НА ПРИМЕРЕ МЕТОДОЛОГИИ)

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

В работе представлен оригинальный 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 методологии.

Ключевые слова: терминологическая структура, теория графов, методология

1. Введение

В соответствии с определением, приведенным в [12], *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 соответствующем *профессиональном языке* [11], 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*

¹ Губанов Дмитрий Алексеевич, к.т.н., с.н.с. (dmitry.a.g@gmail.com)

² Макаренко Андрей Викторович, к.т.н., с.н.с. (avm.science@mail.ru)

³ Новиков Дмитрий Александрович, член-корр. РАН, зам. директора (novikov@ipu.ru)

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, 6, 7, 18, 19]), нас интересуют не общие качественные закономерности формирования и развития терминологии определенной отрасли знания, а количественные свойства ее текущего «среза». 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 качественных вопроса. Первый вопрос – насколько «хорошо» то или иное множество терминов характеризует заданную предметную область, и как построить наиболее адекватное множество терминов. Второй вопрос – как эти термины связаны друг с другом, можно ли, например, «упорядочить» их по общности, важности, встречаемости в различных сочетаниях и т.д., и как это сделать.

Существуют две крайности в ответе на сформулированный вопрос «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, то есть достигается при применении *автоматизированного экспертного подхода* (компьютерные методы осуществляют 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 [34]. 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. Исходно, он состоит из набора статей для определяемых терминов. Каждая из статей имеет структуру, приведённую на рисунке 1а.

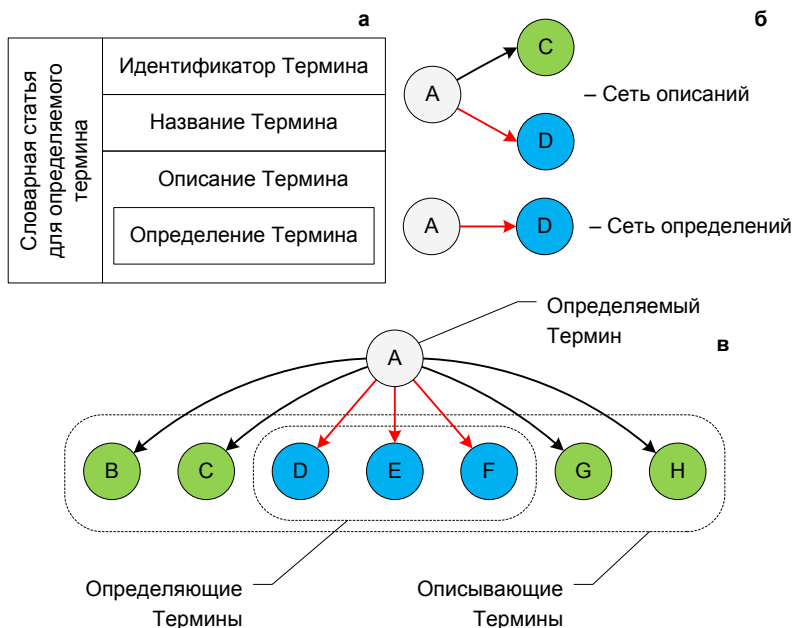


Рис. 1. Описание формата и элементов тезауруса терминологической структуры предметной области

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).

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 «статье» 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 [11] (раздел 4).

2. Подходы к разработке и анализу тезаурусов

2.1. Общие положения

Исследователями уже достаточно давно разрабатываются методики по построению и использованию тезаурусов [22, 23]. На настоящий момент выработан ряд стандартов, регламентирующих формат представления тезаурусов и предоставляющих рекомендации по их созданию [см. напр. 24]. Много усилий предпринимается по созданию методов автоматического построения тезауруса, основанных на уже ставшей традиционной обработке текстов [25, 26, 27], на

извлечении веб-структур (одном из направлений Web Mining) [28], а также на анализе онлайн-энциклопедий [29].

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

– Добров Б.В., Лукашевич Н.В. Научно-исследовательский вычислительный центр МГУ им. М.В.Ломоносова; АНО Центр информационных исследований. В частности ими разработан общественно-политический тезаурус.

– Филиппович Ю.Н., МГТУ им. Н.Э.Баумана.

– Боровикова О.И., Загорюлько Ю.А., Кононенко И.С. Институт систем информатики имени А.П. Ершова СО РАН, г. Новосибирск.

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») [30]. 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» [24]. Разработано семейство формальных языков SKOS для представления тезаурусов, соответствующих парадигме SemanticWeb [31]. 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. Методика составления и анализа тезаурусов

Общая методика по разработке и анализу тезауруса состоит из следующих шагов [22, 23, 32, 33].

(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, напр. 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, напр. 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 вручную просмотреть источник для выбора терминов, которые могут быть полезны непосредственно или в качестве ссылок на другие термины. Также можно использовать программное обеспечение для автоматизации

зированной «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» [23], 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 терминов⁴ (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)

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.

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.

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 получилась достаточно общая методика, схема которой изображена на рисунке 2.

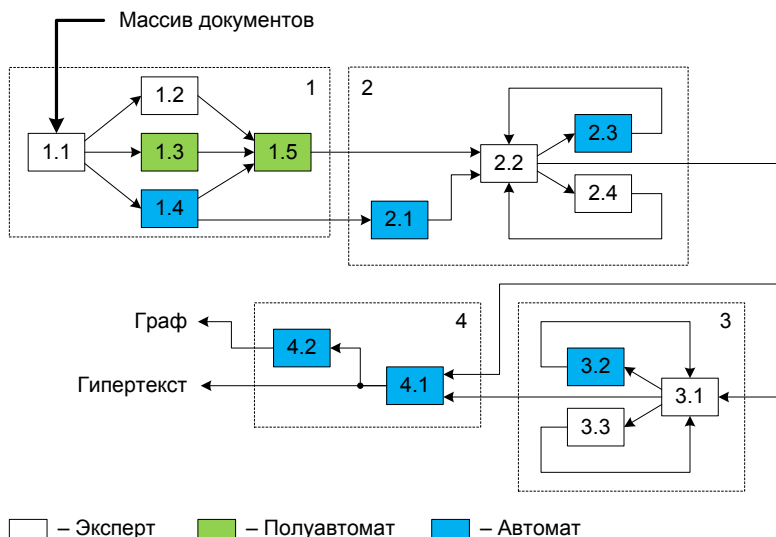


Рис. 2. Методика разработки и анализа терминологической структуры предметной области

Как видно из рисунка 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 необщепотребительных терминов, принадлежащих предметной области;
- 1.3. Анализ структуры библиографических ссылок;
- 1.4. Выделение множества необщепотребительных 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 лингвисти-

ка [35], математическая статистика [36], теория графов [37], программирование [38].

Анализ существующих подходов (см., например, [39]) 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, 4]) 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 метрические 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.

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 ($0 \leq S_i \leq 1$):

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

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 ($0 \leq C_i \leq 1$):

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

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.

4. 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 [11],

в которой приведено наиболее полное, на сегодняшний день, изложение содержания общей методологии.

Операция 1.2. Выделение множества необщепотребительных терминов. Если проанализировать совокупность используемых в упомянутой монографии необщепотребительных терминов, то, во-первых, 200 первых по числу упоминаний терминов дают около 90 % суммарного числа упоминаний всех терминов. Во-вторых, оказывается, что словарь [13] 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 рисунке 3 (обозначена $\hat{I}$) является биссектрисой, нижняя «прямая» (I_j) показывает число терминов словаря, входящих в упорядоченное по убыванию встречаемости в [11] (Q_j) множество терминов методологии. Другими словами, различие между этими прямыми показывает, насколько словарь не соответствует исходному корпусу текстов.

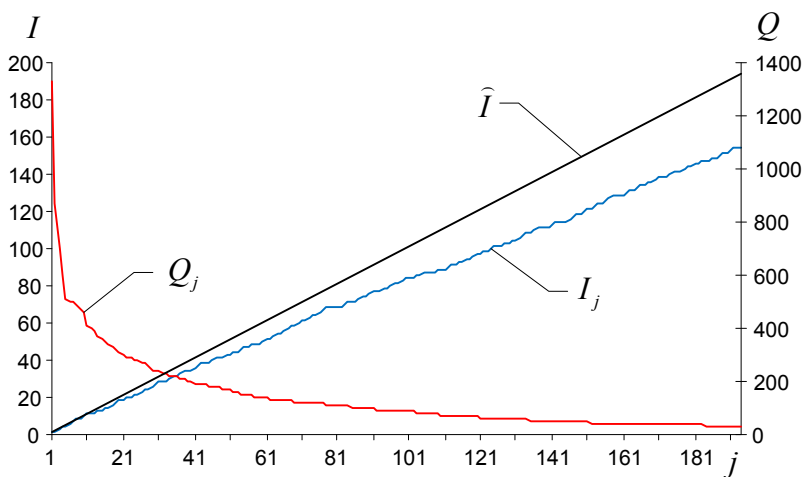
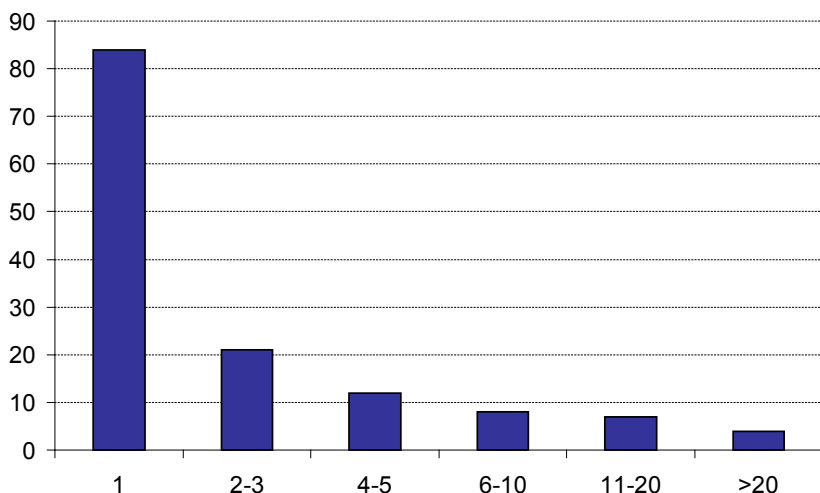


Рис. 3. Распределение терминов [11] по числу упоминаний и соответствие со словарем [13]

Кривая Q_j , в диапазоне $j = \overline{1, 194}$, с достаточно высокой степенью точности, аппроксимируется функцией (средняя относительная ошибка $\sim 4\%$): $\tilde{Q}[j] = \frac{256.605 - 249.079\sqrt{j}}{0.111117 - 0.105957j} - 126.143$. Необходи-

мо 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 [13] типична для любой предметной области (см. описание общих закономерностей в [21]). Большинство терминов заимствовано из работ по методологии [1, 12] (39 и 69 упоминаний соответственно), системному анализу [17] (29 упоминаний), а также из классических словарей по философии [6, 20] (соответственно 17 и 36 упоминаний), логике [1] (17 упоминаний), психологии [5] (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. Сводная информация приведена на рисунке 4.



*Рис. 4. Распределение литературных источников [13]
по числу упоминаний*

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 числом относительно часто встречающихся в предметной области необщепотребительных слов является 250-400. В этот диапазон укладываются как результаты автоматического анализа [11] (300 терминов дают 95 % суммарного числа упоминаний всех терминов), так и экспертно составленный словарь [13] – 297 терминов, а также экспертно составленные словарь по педагогике [10] (около 300 терминов) и включенный в [14] глоссарий по теории управления организационными системами (около 300 терминов – см. также <http://www.mtas.ru/Glossary.htm>).

Наиболее часто в [11] встречаются такие термины, как (перечисление в порядке убывания частоты – см. убывающую кривую, построенную по вспомогательной (правой) оси, на рисунке 3, а также

In, *ClosIn* и *PageRank*. Первые 30 терминов (упорядоченных по убыванию значимости) таковы (ср. с (1)):

- деятельность, процесс, объект, форма, система, цель, явление, метод, знания, теория, действие, результат, организация, наука,
- (2) 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)):

- деятельность, теория, метод, наука, моделирование, проект, системный анализ, проблема, эксперимент, прогнозирование, экспертные оценки, рефлексия, метод сценариев, методология, средство, научная деятельность, семиотика, подход, опытная работа,
- (3) 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 рисунке 5 приведено соответствие между значениями параметров «Значимость» и «Сложность» для сети описаний.

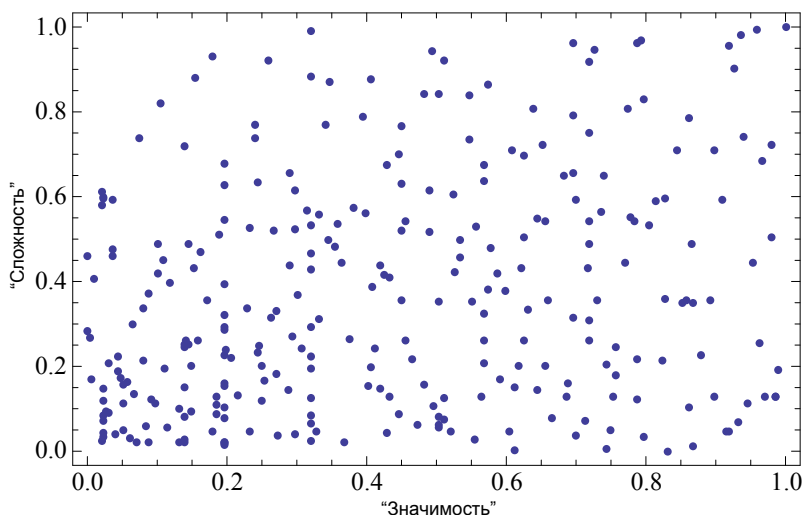


Рис. 5. Соответствие между значениями параметров «Значимость» и «Сложность» для сети описаний

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

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

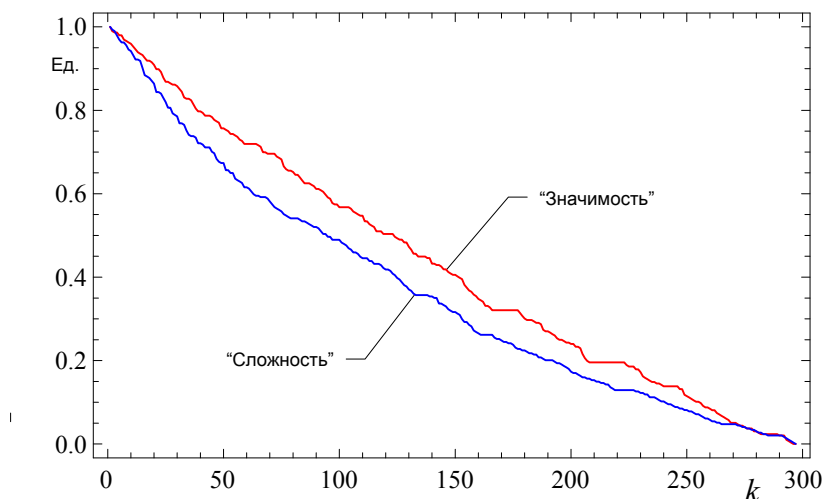


Рис. 6. Графики величин «Значимость» и «Сложность», упорядоченных по убыванию, для сети описаний

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)):

- деятельность, процесс, явление, знания, результат, действие,
 (4) 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)):

- 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 рисунке 7 для сети определений приведено соответствие между значениями параметров «Значимость» и «Сложность».

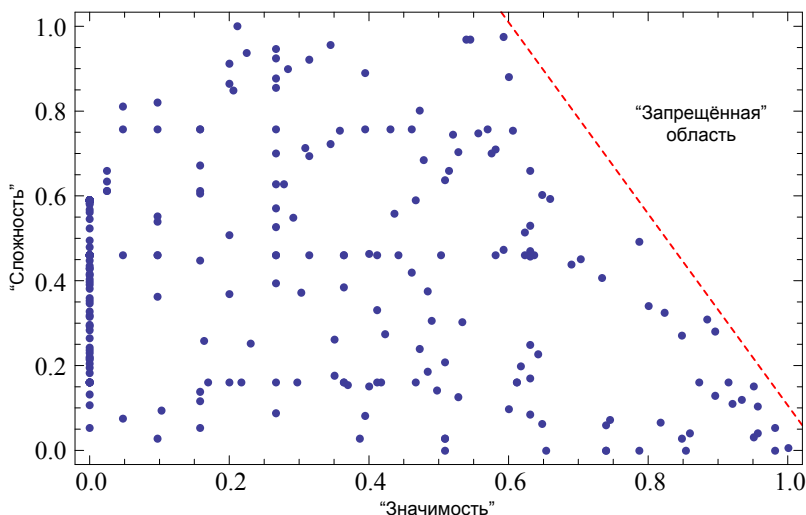


Рис. 7. Соответствие между значениями параметров «Значимость» и «Сложность» для сети определений

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 рисунке 8 для сети определений приведены графики величин «Значимость» и «Сложность», упорядоченных по убыванию.

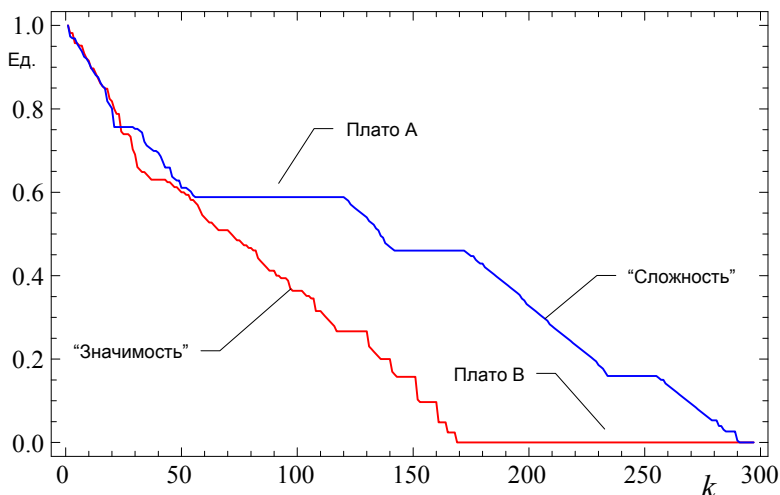


Рис. 8. Графики величин «Значимость» и «Сложность», упорядоченных по убыванию, для сети определений

Из рисунка 8 видно, что имеется значительное число терминов, обладающих высокой одинаковой сложностью (плато А). При этом почти половина терминов имеет минимальную значимость (плато В).

Также существует (коэффициент корреляции Спирмена равен 0.659) статистически значимая положительная ранговая корреляция 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 (то есть 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 (коэффициент корреляции Спирмена равен 0,731) статистически значимая положительная ранговая корреляция между сложностью термина и его максимальным уровнем. Также существует (коэффициент корреляции Спирмена равен –0,480) статистически значимая отрицательная ранговая корреляция между значимостью термина и его максимальным уровнем. Кроме того, следует отметить (коэффициент корреляции Спирмена равен 0,993) статистически значимую ранговую корреляцию 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 [13] (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.

Заключение

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, [15, 16, 36]): 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 [40].

Перспективным представляется возможный будущий анализ терминологической структуры различных предметных областей, что, конечно, потребует от исследователей определенных усилий (см. выше неавтоматизируемые, экспертные компоненты методики на рисунке 2), но в значительной степени облегчается наличием уже созданных (и существующих в электронном виде) семантических сетей и онтологий во многих отраслях научного знания.

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.

Литература

1 Блох М.Я. Теоретические основы грамматики. – М.: Высшая школа, 2004.

- 2 Голованова Е.И. Введение в когнитивное терминоведение. – М.: Флинта, 2011.
- 3 Ивин А.А. Логика. – М.: Знание, 1998.
- 4 Кондаков Н.И. Логический словарь-справочник. – М.: Наука, 1975.
- 5 Краткий психологический словарь / Сост. Л.А. Карпенко. Под общ. ред. А.В. Петровского, М.Г. Ярошевского. – М.: Политиздат, 1985.
- 6 Лейчик В.М. Терминоведение. – М.: ЛКИ, 2007.
- 7 Лотте Д.С. Основы построения научно-технической терминологии / Вопросы теории и методики. – М.: Изд-во академии наук СССР, 1968.
- 8 Новая философская энциклопедия: В 4 тт. / Под редакцией В.С. Стёпина. – М.: Мысль, 2000-2001.
- 9 Новиков А.М. Основания педагогики. 2-е изд. – М.: Эгвес, 2011.
- 10 Новиков А.М. Педагогика: словарь системы основных понятий. – М.: Издательский центр ИЭТ, 2013.
- 11 Новиков А.М., Новиков Д.А. Методология. – М.: Синтег, 2007.
- 12 Новиков А.М., Новиков Д.А. Методология научного исследования. – М.: Либроком, 2010.
- 13 Новиков А.М., Новиков Д.А. Методология: словарь системы основных понятий. – М.: Либроком, 2013.
- 14 Новиков Д.А. Теория управления организационными системами 3-е изд. испр. и дополн. – М.: Физматлит, 2012.
- 15 Новиков Д.А., Новочадов В.В. Статистические методы в медико-биологическом эксперименте (типовые случаи). – Волгоград: Издательство ВолГМУ, 2005.
- 16 Орлов А.И. Эконометрика. – М.: Экзамен, 2004.
- 17 Перегудов Ф.И., Тарасенко Ф.П. Введение в системный анализ. – М.: Высшая школа, 1989.
- 18 Реформатский А.А. Терминология. Введение в языкознание. – М.: Учпедгиз, 1959.
- 19 Суперанская А.В., Подольская Н.В., Васильева Н.В. Общая терминология. Вопросы теории. Изд. 4-е. – М.: Изд-во ЛКИ, 2007.
- 20 Философский энциклопедический словарь. – М.: Сов. Энциклопедия, 1983.

21 Яблонский А.И. Модели и методы исследования науки. – М.: Эдиториал УРСС, 2001.

22 Aitchison J., Gilchrist A., Bawden D. Thesaurus Construction And Use: A Practical Manual. - Aslib, 2000. 218 p.

23 Sager J.C. A practical course in terminology processing. - J. Benjamins Pub. Co., 1990. 254 p.

24 ISO 25964-1,
http://www.iso.org/iso/catalogue_detail.htm?csnumber=53657

25 Manning C.D., Schütze H. Foundations of statistical natural language processing. MIT, 1999.

26 Crouch C.J. A cluster based approach to thesaurus construction. Proc. of the ACM SIGIR, pages 309–320, 1988.

27 Schütze H., Pedersen J.O. A cooccurrence-based thesaurus and two applications to information retrieval. International Journal of Information Processing and Management, 33. 1997. P. 307–318.

28 Chen Z., Liu S., Wenyin L., Pu G., Ma W.Y. Building a web thesaurus from web link structure. Proc. of the ACM SIGIR. 2003. P. 48–55.

29 Strube M., Ponzetto S. WikiRelate! Computing semantic relatedness using Wikipedia. In Proc. of National Conference on Artificial Intelligence (AAAI'06). 2006. P. 1419–1424.

30 Нгуен М.Х., Аджиев А.С. Описание и использование тезаурусов в информационных системах, подходы и реализация. – URL: <http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2004/part1/NA>

31 SKOS Simple Knowledge Organization System. – URL: <http://www.w3.org/2004/02/skos/>

32 Handbook of Terminology Management: Basic Aspects of Terminology Management. Sue Ellen Wright, Gerhard Budin. John Benjamins Pub Co, 1997. 370 p.;

33 Grefenstette G. Explorations in Automatic Thesaurus Discovery. 1994

34 Тезаурус. -URL: <http://ru.wikipedia.org/wiki/Тезаурус>

35 Bird S., Klein E. Loper E. Natural Language Processing with Python. O'Reilly Media, 2009.

36 Кобзарь А.И. Прикладная математическая статистика. – М.: Физматлит, 2006.

37 Харари Ф. Теория графов. – М.: Мир, 1973.

38 Кормен Т.Х. и др. Алгоритмы: построение и анализ. – М. Вильямс, 2006.

39 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. Т.6, В.3 (2012), С. 22–31.

40 Witten I.H., Frank E., Hall M.A. Data Mining: Practical Machine Learning Tools and Techniques (Third Edition), Morgan Kaufmann, 2011.