УДК 51-7

МЕТОДЫ ИЗВЛЕЧЕНИЯ И АНАЛИЗА ТЕРМИНОЛОГИЧЕСКИХ СТРУКТУР СМЕЖНЫХ ПРЕДМЕТНЫХ ОБЛАСТЕЙ (НА ПРИМЕРЕ МЕТОДОЛОГИИ)

Д.А. Губанов[1], Д.А. Новиков[2]

Институт проблем управления имени В.А. Трапезникова РАН, Москва, Россия
[1]*dmitry.a.g@gmail.com*, [2]*novikov@ipu.ru*

Аннотация

В работе предложен автоматизированный экспертный подход к извлечению терминологических структур из монографий. Постулируется, что содержание предметной области целиком отражается в некоторой монографии, это позволяет извлекать и анализировать терминологические структуры без использования заранее подготовленного корпуса специализированных текстов предметной области. В рамках подхода рассмотрена процедура итеративного извлечения и анализа терминологических структур, регулирующая работу эксперта с монографией и позволяющая ему контролировать процесс. Предложены методы сравнительного анализа терминологических структур двух смежных предметных областей. Возможности разработанного подхода продемонстрированы на примере общей методологии и методологии комплексной деятельности.

Ключевые слова: *терминологическая структура, теория графов, общая методология, методология комплексной деятельности.*

Цитирование: *Губанов, Д.А.* Методы извлечения и анализа терминологических структур смежных предметных областей (на примере методологии) / Д.А Губанов, Д.А. Новиков // Онтология проектирования. – 2018. – Т. 8, №3(29). – С.347-365. – DOI: 10.18287/2223-9537-2018-8-3-347-365.

Введение

В соответствии с определением, приведенным в [1], *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 (ПрО)*, то есть всех явлений и процессов, описываемых данной теорией.

Изложение научных результатов, полученных в той или иной ПрО, ведется, как правило, на соответствующем профессиональном языке, использующем, помимо общеупотребительной, свою специальную терминологию. То есть в каждой ПрО можно условно поставить в соответствие множество терминов, характеризующих эту ПрО и используемых в ней [2]. Совокупность этих терминов и связей между ними будем называть *терминологической структурой* ПрО. Постулируем, что содержание ПрО целиком отражается в некоторой *монографии*[1], поэтому исходными данными для анализа терминологической структуры является текст этой монографии.

В настоящей работе рассматривается задача автоматизированного извлечения терминологической структуры из текста соответствующей монографии, а также формулируется задача сравнительного анализа терминологических структур различных ПрО.

[1] Согласно *ГОСТ 7.60-2003,* монография - «Научное или научно-популярное издание, содержащее полное и всестороннее исследование одной проблемы или темы …».

В разделе 1 рассматриваются методы извлечения терминологической структуры ПрО, в разделе 2 – методы сравнительного анализа терминологических структур смежных ПрО, в разделе 3 приводится сценарий организации процесса извлечения терминов из текста. В разделе 4 приводятся примеры анализа терминологических структур таких ПрО как *общая методология* (ОМ) – см. монографию [1] – и *методология комплексной деятельности* (МКД) – см. монографию [3], а также приводится их сравнительный анализ.

Методология - учение об организации деятельности. В [1] изложена ОМ (основания ОМ, характеристики деятельности, ее логическая и временн*а*я структуры), а также рассмотрены методология научной, практической, художественной, учебной и игровой деятельности. Монография [3] посвящена изложению МКД, развивающей ОМ на случай любой сложной человеческой деятельности (см. [4]). Значительное внимание в МКД уделяется организации и управлению, неопределённости, а также жизненным циклам (ЖЦ) деятельности, её субъектов, предметов, ресурсов, знаний и технологий.

1 Методы извлечения терминологической структуры ПрО

Используемый подход предполагает последовательное решение следующих взаимосвязанных задач: извлечение терминов и извлечение связей между терминами.

1.1 Извлечение терминов

Пусть имеется *монография* (один документ), результатом должен быть ранжированный список терминов. Общая последовательность этапов имеет следующий вид (см., например, подобные примеры последовательности в [5, 6]):

1) предварительная обработка текста,
2) выявление терминов-кандидатов (или кандидатов в термины),
3) вычисление признаков терминов-кандидатов,
4) отбор кандидатов.

Рассмотрим подробнее эти этапы.

Этап 1. *Предварительная обработка текста* может быть представлена как следующая последовательность действий.

- Преобразование исходного файла в текстовый формат, фильтрация артефактов преобразования и неинформативных элементов.
- Токенизация полученного текста (разбиение текста на слова).
- Лемматизация слов или стемминг.

Этап 2. *Выявление терминов-кандидатов.* На этом этапе формируется множество кандидатов в термины (потенциальных терминов).

- *Выявление n-грамм* (терминов-кандидатов).
- *Фильтрация терминов-кандидатов.* Применяются экспертные знания о том, какие кандидаты в термины заведомо не являются терминами. На практике полезными являются следующие виды фильтров: лингвистическая фильтрация (например, рассматриваются только кандидаты-словосочетания с существительным в роли главного слова); фильтрация кандидатов по частоте встречаемости; фильтрация по содержанию в составе кандидата стоп-слов (по стоп-слову из экспертного списка, по стоп-слову в начале или в конце кандидата стоп-слова (например, «и»), по символу пунктуации (например, «,»)); фильтрация по длине кандидата (по числу слов в последовательности); фильтрация на основе регулярных выражений и др.

Этап 3. *Расчет значений признаков терминов-кандидатов*. На этом этапе формируются и рассчитываются значения признаков-кандидатов, определяющих их вероятность быть термином.

- *Формирование множества признаков терминов-кандидатов*. Для терминов-кандидатов определяется список основных и вспомогательных признаков (таким образом каждый кандидат может быть представлен в виде вектора признаков). Основные признаки характеризуют терминологичность кандидата, т.е. силу связи кандидата с ПрО. В массе своей это признаки, которые основываются на статистиках вхождений слова в сегменты документа (или документ, или корпус документов) и используют разного рода сопутствующие эвристики («навешивающие» вознаграждения и штрафы на те или иные статистики). Примером является абсолютная и относительная частота слова в документе. Вспомогательные же признаки косвенно определяют значимость кандидата. Например, значимость кандидата выше, если он находится в аннотации документа или в секции ключевых слов (таким образом можно определить множество структурных признаков, характеризующих встречаемость слова в различных логических разделах документа). Также можно определить признаки, характеризующие вхождение кандидата во внешнюю энциклопедию (напр., Википедию, см. [7]), словарь, тезаурус или онтологию; использовать признаки, характеризующие форматирование слова в исходном документе (курсив, жирный шрифт и т.д.); использовать признаки, характеризующие наличие тех или иных частей речи в кандидате.
- *Расчёт значений признаков.* Производится расчет признаков кандидата. Предлагается использовать расчет основных признаков с использованием статистик вхождений кандидата в текст (см. [8, 9]).
- *Расчёт интегрального показателя.* Далее на основе совокупности признаков кандидатов оценивается общая значимость каждого кандидата (оценка того, насколько он «является термином»). Такая оценка может быть вычислена при помощи различных методов: линейной комбинации признаков, алгоритмов голосования или машинного обучения.

Этап 4. *Отбор кандидатов.* Кандидаты ранжируются и отбирается топ (например, первые 300 кандидатов). Затем ранжированный список кандидатов оценивается и фильтруется экспертом в соответствующей ПрО.

Приведём примеры типовых подходов, в которых реализуются те или иные шаги приведённого алгоритма-шаблона.

Пример 1. Подход на основе статистик вхождения:

1) формирование списка 1 и 2-грамм, приведённых в нормальную форму;
2) фильтрация *n*-грамм, содержащих слова из стоп-листа или имеющих размер (в символах) меньше некоторого порога;
3) расчёт частоты *n*-грамм (и мер сопряжённости для биграмм);
4) ранжирование *n*-грамм по убыванию значения статистики и отбор топ-*K* терминов (где $K \approx 10^2$).

Пример 2. Пример подхода на основе морфологических шаблонов (например, <существительное> + <существительное в родительном падеже>, см., например, [10]):

1) разработка морфологических шаблонов;
2) определение кандидатов при помощи шаблонов;
3) расчёт частоты встречаемости кандидатов;
4) ранжирование кандидатов и отбор.

1.2 Извлечение связей между терминами

Для построения терминологической структуры ПрО необходимо установить связи между входящими в неё терминами.

Возможны следующие подходы к извлечению связей между терминами.

На основе лексико-синтаксических шаблонов. Извлечение из корпуса текстов родовидовых отношений с использованием лексико-синтаксических шаблонов (см., например, [12]).

На основе статистик совместной встречаемости. Если термины встречаются совместно, то между ними может существовать связь. Самый простой вариант статистики: чем чаще термины встречаются вместе, тем сильнее между ними связь.

Более сложные варианты основываются *на векторном представлении документов / слов*. Традиционно используемый для выявления связей между словами метод основан на формировании матрицы совместной встречаемости слов в текстах коллекции документов ПрО (на основе матрицы «слова-на-документы»), оценке сходства рассматриваемых терминов и – в конечном итоге - формировании связи между терминами. Среди недостатков этого подхода (разреженность матрицы совместной встречаемости слов, большой размер матрицы и ее зашумленность) наиболее критичным в нашем случае является необходимость работы с коллекцией документов.

Один из современных подходов к векторному представлению слов – «word embeddings» [13] (при этом не требуется использовать большую матрицу «слова-на-документы», в процессе работы используется ближайшая окрестность слов). В рамках этого подхода алгоритмы (например, нейронные сети) прогнозируют появление слова в заданном контексте, и в процессе обучения алгоритма формируется матрица параметров (в том числе векторное представление слов небольшой размерности). Полученное векторное представление слов можно использовать для оценки сходства терминов (см. пример извлечения сети семантических связей из Википедии [14]).

В данной работе сила связи между терминами оценивается исходя из их совместной встречаемости в предложениях исследуемого документа (соответствующей монографии).

2 Методы сравнительного анализа терминологических структур смежных ПрО

Установление соответствий и различий между двумя ПрО и их терминологическими структурами может быть полезно для решения следующих задач:

- анализ содержательного соотношения ПрО;
- согласование терминологических структур;
- обогащение одной терминологической структуры за счёт отсутствующих в ней элементов из другой терминологической структуры;
- интеграция терминологических структур, в процессе которой происходит создание новой терминологической структуры, и т.д.

Для проведения сравнительного анализа терминологических структур предлагаются следующие подходы.

2.1 Статистический подход

Статистический подход основывается на сравнении встречаемости терминов в текстах двух ПрО (здесь и далее предполагается, что для сравнительного анализа терминологических структур важна значимость термина и контекст появления термина).

Статистики отдельного термина:

- сравнение частот вхождений термина в тексты разных ПрО;
- сравнение контекстов использования термина в текстах разных ПрО, расчёт близости/схожести использования термина по контекстам встречаемости.

Статистики терминологической структуры:

- сравнение ранжирования терминов в двух ПрО, получение оценки сходства ранжирований (например, при помощи ранговой корреляции).

- сравнение вероятностных распределений терминов в двух ПрО, получение оценки сходства вероятностных распределений (например, при помощи расстояния Кульбака-Лейблера).

2.2 Структурно-статистический подход

Структурно-статистический подход основывается на сравнении статистик, связанных с отдельными вершинами графа терминологической структуры и/или статистик, характеризующих весь граф терминологической структуры.

Сравнение терминов (вершин графа).

- Сравнение значимости термина в разных терминологических сетях (показатели структурной центральности, показатели влиятельности, см. [15]). Для проведения сравнительного анализа в работе используются следующие виды структурных центральностей: центральность вершины по степени (число инцидентных вершине ребер в графе/сети), центральность вершины по близости (величина обратная сумме расстояний от вершины до всех остальных вершин сети), центральность вершины по посредничеству (оценивает то, как часто вершина появляется на кратчайших путях между парами вершин сети), центральность вершины по PageRank (является результатом случайного блуждания по вершинам сети), а также влиятельность вершины, основанная на акциональной модели [16]. Следует отметить, что рассчитать такую влиятельность для узлов статичных неориентированных сетей можно при помощи имитационного моделирования распространения действий (в данном случае в качестве модели распространения действий выбрана модель независимых каскадов [17]).
- Оценка сходства терминов. Термины (вершины) считаются похожими, если они прямо или косвенно связаны с похожими терминами. Например, можно сравнивать использование термина, рассчитывая расстояние между эго-сетями [16] термина в разных терминологических структурах.

Сравнение терминологических структур (графов).

- Сравнение ранжирований терминов в разных терминологических структурах по убыванию значимости (структурной центральности).
- Сравнение графов и оценка их структурного сходства. Сходство двух графов (близость) можно определить, например, с использованием максимально общего изоморфного подграфа. В частности, если обозначить граф g_1, граф g_2 и максимально общий изоморфный подграф $mcs(g_1, g_2)$, то сходство g_1 и g_2 оценивается следующим образом: $\frac{|mcs(g_1,g_2)|}{\max\{|g_1|,|g_1|\}}$.

Кроме того, расстояние между графами можно рассчитать при помощи минимального количества операций редактирования (вставки, удаления и замены вершин или ребер графа), необходимых для превращения одного графа в другой граф (Graph Edit Distance).

Можно использовать для сравнения и оценки сходства векторное представление графов. Примерами признаков могут быть типовые сетевые структуры, в частности, множество эго-сетей заданных экспертом понятий или множество всех путей заданной длины, содержащихся в графах (или наиболее вероятные / значимые маршруты). Далее рассчитываются значения признаков графов, и вычисляется расстояние между векторами графов.

3 Процесс извлечения терминов

Извлечение терминов ПрО из монографии производится в полуавтоматическом режиме, в этот процесс вовлечены эксперт в рассматриваемой ПрО и разработчик (или соответствующая программа, обладающая развитым пользовательским интерфейсом). Сценарий взаимодействия эксперта и разработчика представлен на рисунке 1.

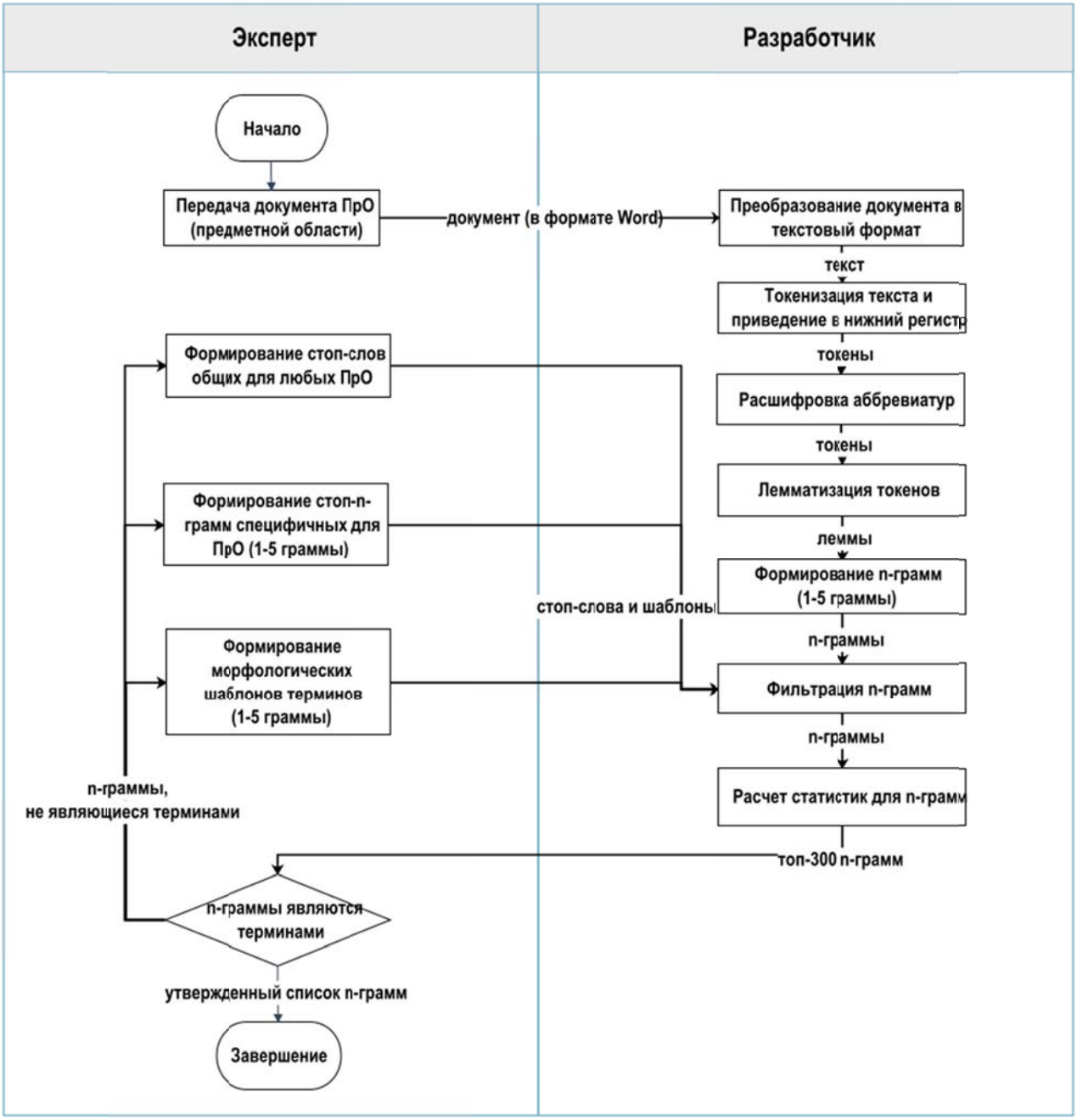

Рисунок 1 – Сценарий взаимодействия эксперта и разработчика

Приведём некоторые пояснения к рисунку 1.

1. Расшифровка аббревиатур. Каждый токен в последовательности, принадлежащий экспертному списку аббревиатур, заменяется на соответствующие токены из расшифровки. Например, аббревиатура «СЭД» заменяется на «структурный», «элемент», «деятельности».

2. Лемматизация токенов. Производится при помощи размеченного корпуса русского языка OpenCorpora (opencorpora.org). Часть речи слова и его начальная форма определяются исходя из частоты встречаемости (слово + часть речи) в корпусе OpenCorpora.

3. Формирование n-грамм. Из ранее полученной последовательности лемм текста формируются все допустимые последовательности из *n* лемм.

4. Фильтрация n-грамм. Используемые правила фильтрации:

- частота встречаемости *n*-граммы в тексте не менее 5;
- длина первого и последнего слова в *n*-грамме больше 1;
- *n*-граммы не должны содержать символы пунктуации;
- *n*-граммы не должны содержать числа;
- *n*-граммы не должны содержать в себе общеупотребительные слова (возможно дополненных экспертом), а также не должны заканчиваться или начинаться предлогами или союзами;
- *n*-граммы не должны входить в составленный экспертом список стоп-n-грамм (эти *n*-граммы не являются терминами ПрО, их список может быть сформирован заранее, а затем дополнен экспертом при просмотре результатов извлечения терминов из текста);
- *n*-граммы должны соответствовать определённому морфологическому шаблону, например:

	1 слово	2 слово	3 слово	4 слово	5 слово
1-грамма	Сущ/Прил/Прич/?				
2-грамма	Сущ/Прил/Прич/?	Сущ/?			
3-грамма	Сущ/Прил/Прич/?		Сущ/?		
4-грамма	Сущ/Прил/Прич/?			Сущ/?	
5-грамма	Сущ/Прил/Прич/?				Сущ/?

4 Анализ терминологической структуры

4.1 Анализ терминологической структуры МКД

В качестве исходного корпуса текста используется монография [3] .

4.1.1 Выделение терминов–униграмм

Выделение терминов из текста проводилось в рамках предложенного подхода (раздел 1) согласно рассмотренному в разделе 3 сценарию. Всего получено 134 термина, наиболее часто в тексте встречаются следующие термины (топ-30 в порядке убывания частоты):
деятельность, элемент, комплексный, технология, модель, система, субъект, ресурс, результат, неопределённость, управление, структура, цикл, жизненный, операция, предмет, организация, цель, событие, элементарный, спрос, причинно-следственный, вид, логический, методология, потребность, характеристика, процесс, работа, связь. Распределение терминов МКД представлено на рисунке 2.

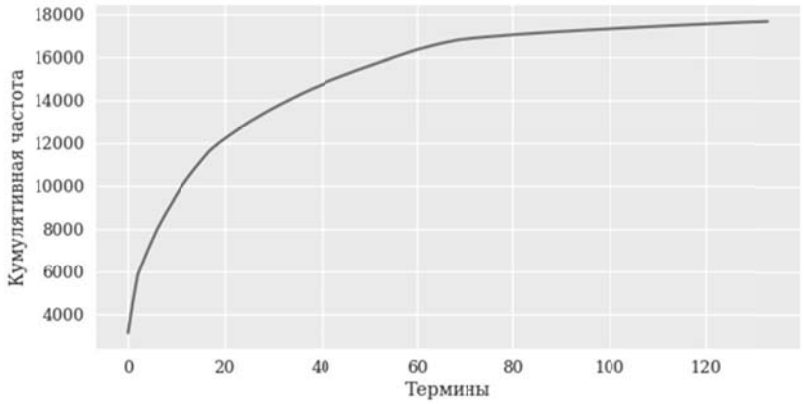

Рисунок 2 – Распределение терминов МКД по убыванию частоты

Основные термины МКД в виде облака тегов представлены на рисунке 3.

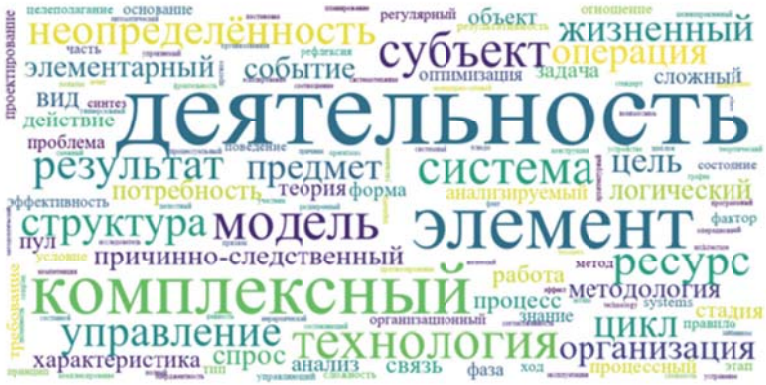

Рисунок 3 – Облако терминов МКД

4.1.2 Выделение терминов – n-грамм (n > 1)

Ранжирование *n*-грамм производилось при помощи различных методов: частота (прямой подсчет количества n-грамм), поточечная взаимная информация, *t*-тест, χ^2-тест, отношение функций правдоподобия (см. [8]). Оказалось, что согласно экспертной оценке самый простой способ (ранжирование по частоте) – наилучший (этот вывод подтверждается результатами другого экспериментального исследования [18]).

Биграммы МКД (топ-30):

комплексный деятельность, элемент деятельность, структурный элемент, ЖЦ, организационно-технический система, элементарный операция, пул ресурс, информационный модель, логический структура, процессный модель, причинно-следственный структура, новый технология, причинно-следственный модель, внешний среда, событие неопределённость, результат деятельность, создание технология, нижестоящий операция, нижестоящий элемент, предмет деятельность, субъект деятельность, формирование спрос, цикл ресурс, регулярный деятельность, технология деятельность, systems engineering, достижение цель, получение результат, истинный неопределённость, репликативный деятельность.

3-граммы МКД (топ-30):

структурный элемент деятельность, элемент комплексный деятельность, нижестоящий структурный элемент, анализируемый структурный элемент, методология комплексный деятельность, организация и управление, вышестоящий структурный элемент, модель комплексный деятельность, структура комплексный деятельность, создание новый технология, цикл комплексный деятельность, реализация ЖЦ, выполнение комплексный деятельность, стадия ЖЦ, ЖЦ ресурс, результативность и эффективность, технология комплексный деятельность, реализация комплексный деятельность, модель структурный элемент, субъект комплексный деятельность, реакция на неопределённость, результат комплексный деятельность, целеполагание и создание, управление комплексный деятельность, ЖЦ пул, процедура формирование спрос, неопределённость комплексный деятельность, предмет комплексный деятельность, деятельность в целое, цикл пул ресурс.

4-граммы МКД (топ-30):

нижестоящий структурный элемент деятельность, анализируемый структурный элемент деятельность, вышестоящий структурный элемент деятельность, ЖЦ комплексный деятельность, субъект анализируемый структурный элемент, модель структурный элемент деятельность, логический структура комплексный деятельность, реализация ЖЦ ресурс, субъект нижестоящий структурный элемент, субъект вышестоящий структурный элемент, ЖЦ пул ресурс, деятельность и элементарный операция, логический и причинно-следственный структура, структура структурный элемент деятельность, выполнение структурный элемент деятельность, порождение элемент комплексный деятельность, процессный модель структурный элемент, целеполагание и создание технология, креативный структурный элемент деятельность, модель элемент комплексный деятельность, ЖЦ структурный элемент, комплексный деятельность в целое, спрос и осознание потребность, фиксация спрос и осознание, цикл структурный элемент деятельность, цикл элемент комплексный деятельность, действие и получение результат, оптимизация выполнение комплексный деятельность, субъект структурный элемент деятельность, цепочка реакция на неопределённость.

5-граммы МКД (топ-30):

субъект анализируемый структурный элемент деятельность; субъект нижестоящий структурный элемент деятельность; субъект вышестоящий структурный элемент деятельность; элемент деятельность и элементарный операция; процессный модель структурный элемент деятельность; результативность и эффективность комплексный деятельность; фиксация спрос и осознание потребность; ЖЦ элемент комплексный деятельность; ЖЦ структурный элемент деятельность; требование к методология комплексный деятельность; организация и управление комплексный деятельность; комплексный деятельность с известный технология; комплексный деятельность и организационно-технический система; выполнение действие и получение результат; логический структура структурный элемент деятельность; элемент деятельность и нижестоящий операция; запрос, получение и организация; фаза , стадия и этап; целеполагание и создание новый технология; деятельность, структурный элемент деятельность; задача оптимизация выполнение комплексный деятельность; внутренний упорядоченность , согласованность взаимодействие; стадия ЖЦ комплексный деятельность; порождение новый элемент комплексный деятельность; текущий содержание конкретный агрегат оборудование; структура анализируемый структурный элемент деятельность; модель анализируемый структурный элемент деятельность; анализ , синтез, конкретизация; синтез , конкретизация , регулирование; утверждение о состав комплексный деятельность.

Дальнейший анализ проводится только для терминов, являющихся униграммами.

4.1.3 Построение сети терминов МКД

На основе совместной встречаемости терминов (в пределах предложений текста) построен взвешенный граф терминов, в котором вершинами являются термины, а ребрами – связи между ними, вес ребра равен числу совместных появлений соответствующей пары терминов.

На рисунке 4 показана наибольшая компонента связности графа терминов МКД с ребрами, вес которых больше 50. На дугах – частота совместного появления терминов в предложениях текста.

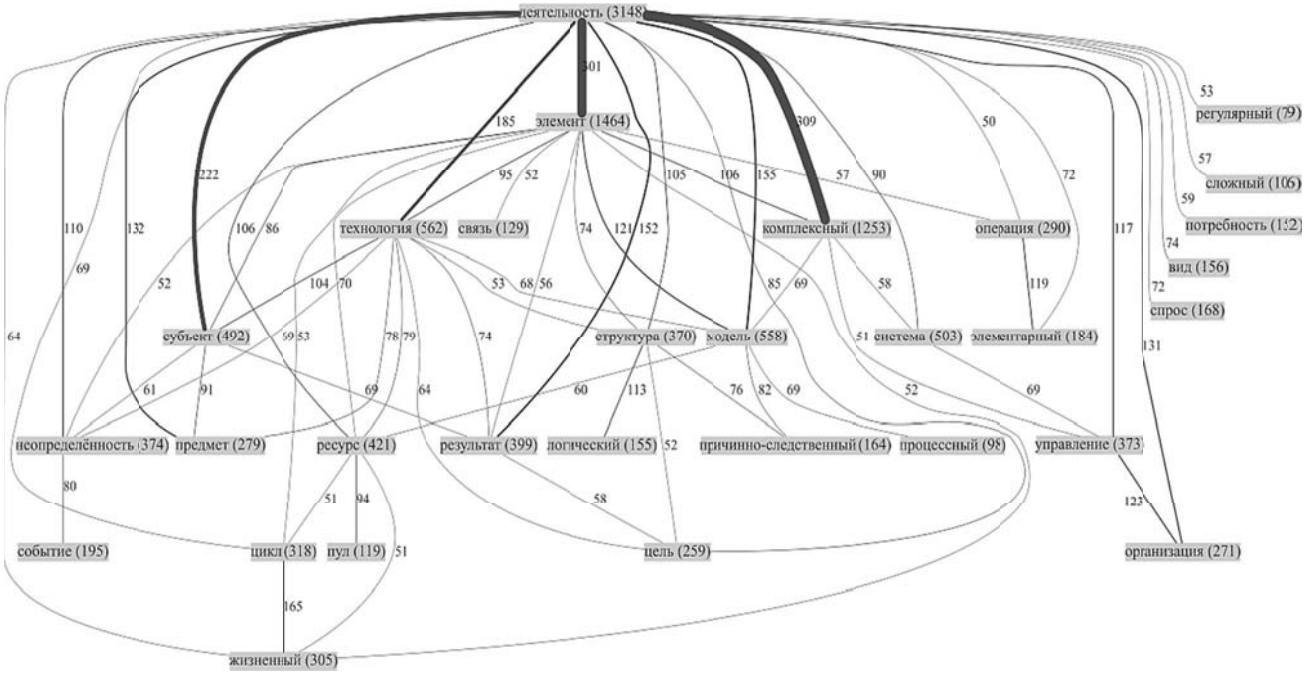

Рисунок 4 – Граф терминов МКД

4.1.4 Анализ структурных центральностей терминов

Имея сеть терминов, можно для каждого термина рассчитать его структурную значимость и таким образом определить наиболее важные термины в интерпретации, зависящей от используемого показателя (см. таблицу 1). Для оценки структурной значимости терминов используются следующие показатели: центральность по степени, центральность по PageRank, центральность по близости, центральность по посредничеству, а также влиятельность (влиятельность по акциональной модели).

Таблица 1 – Топ-10 терминов МКД по каждому из показателей

№	Частота	Степень	Близость	Посредничество	Pagerank	Влиятельность
1	деятельность	деятельность	деятельность	деятельность	деятельность	деятельность
2	элемент	элемент	элемент	система	элемент	элемент
3	комплексный	система	комплексный	элемент	технология	технология
4	технология	субъект	субъект	технология	модель	модель
5	модель	структура	технология	модель	субъект	комплексный
6	система	модель	модель	результат	комплексный	субъект
7	субъект	технология	результат	управление	результат	результат
8	ресурс	ресурс	организация	неопределённость	ресурс	ресурс
9	результат	комплексный	управление	ресурс	структура	система
10	неопределённость	вид	предмет	структура	система	структура

На рисунке 5 представлены значения коэффициентов корреляции Спирмена для значений показателей терминов МКД. Высокие коэффициенты ранговой корреляции Спирмена (> 0,9) имеют место между пятью переменными – частота, степень, близость, PageRank и влиятельность. Исходя из этого, можно считать, что для оценки структурной значимости терминов достаточно использовать частоту встречаемости терминов и центральность по посредничеству.

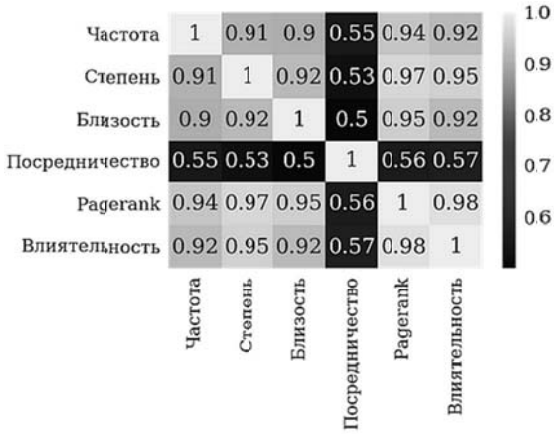

Рисунок 5 – Коэффициенты корреляции Спирмена для МКД

4.1.5 Выделение кластеров терминов МКД в сети

Для выделения кластеров терминов в взвешенном графе использованы методы выявления сообществ теории социально-сетевого анализа. Качество кластеризации обычно оценивается при помощи показателя модулярности, принимающего значения от –1 до +1 (больше – лучше). Наилучшие результаты показал метод спинового стекла [19], качество кластеризации составило 0,124.

При помощи метода спинового стекла получено четыре кластера терминов:

1) ресурс, цикл, жизненный, пул, стадия, форма, фаза, проектирование, этап, рефлексия, эксплуатация, планирование, операционный, график, взаимосвязь, мера, календарно-сетевой, стоимость, увольнение, компетенция;

2) элемент, модель, структура, операция, элементарный, причинно-следственный, вид, логический, связь, анализируемый, процессный, организационный, тип, принцип, управляющий, агент, механизм, иерархический, комплексирование, организованная, полный, фрактальность, конструкция, исследователь;

3) технология, субъект, результат, неопределённость, цель, событие, спрос, потребность, характеристика, действие, задача, требование, оптимизация, эффективность, условие, регулярный, правило, отношение, фактор, ход, синтез, целеполагание, результативность, эффект, параметр, полезность, прогноз, устранение, эскалирование, участник, повторение, блюдо, факт, постановка, желаемый, соотношение, прогнозирование;

4) деятельность, комплексный, система, управление, предмет, организация, методология, процесс, работа, анализ, объект, теория, сложный, знание, проблема, основание, часть, systems, метод, состояние, поведение, сложность, процессуальный, согласованность, причина, управляемый, признак, расширенный, системный, системотехника, стандарт, complex, архитектурный, целенаправленный, программный, ценность, логика, теоретический, смежный, эмерджентность, универсальный, шаблон, architecture, operations, business, составляющий, устройство, составной, попытка, автоматический, целостный, методологический.

Такие кластеры имеют хорошую содержательную интерпретацию и соотносятся со структурой МКД: кластер 1 включает, в основном, термины, связанные с причинно-следственной (временной) структурой комплексной деятельности (КД) и с жц её элементов; кластер 2 – с логической структурой КД, кластер 3 – с процессной структурой и технологией

КД, кластер 4 – с организацией и управлением КД и с организационно-техническими системами.

4.2 Анализ терминологической структуры ОМ

В качестве исходного корпуса текста ОМ используется монография [1] .

4.2.1 Выделение терминов

Выделение терминов (только униграмм) из текста проводилось согласно сценарию, рассмотренному в разделе 3. Всего получено 150 терминов, наиболее часто в тексте встречаются следующие термины (топ-30 в порядке убывания частоты):

деятельность, научный, система, метод, исследование, процесс, проект, результат, знание, игра, учебный, человек, наука, работа, цель, обучение, художественный, проблема, теория, модель, методология, организация, форма, объект, средство, задача, вид, решение, общий, развитие.

Распределение терминов ОМ показано на рисунке 6, а основные термины ОМ в виде облака тегов представлены на рисунке 7.

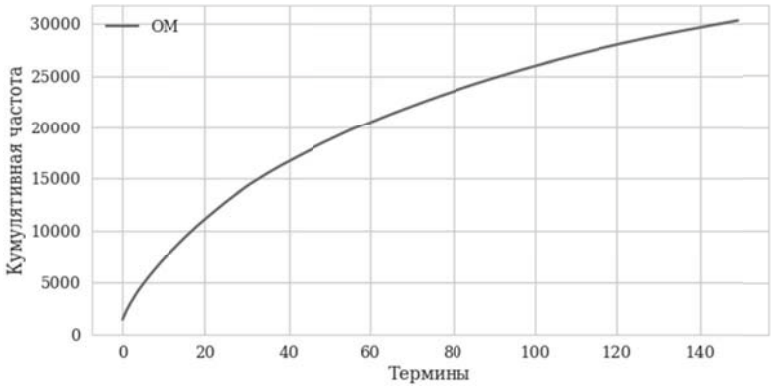

Рисунок 6 – Распределение терминов ОМ по убыванию частоты

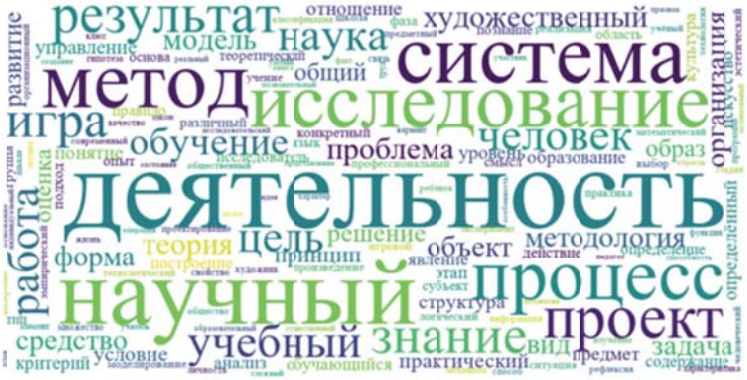

Рисунок 7 – Облако терминов ОМ

4.2.2 Построение сети терминов ОМ

На основе совместной встречаемости терминов ОМ в предложениях текста [1] построен взвешенный граф терминов, в котором вершинами являются термины, а рёбрами – связи между ними, вес ребра равен числу совместных появлений соответствующей пары терминов. На рисунке 8 показана наибольшая компонента связности графа терминов ОМ с рёбрами, вес которых больше или равен 50. На дугах – частота совместного появления терминов в предложениях текста.

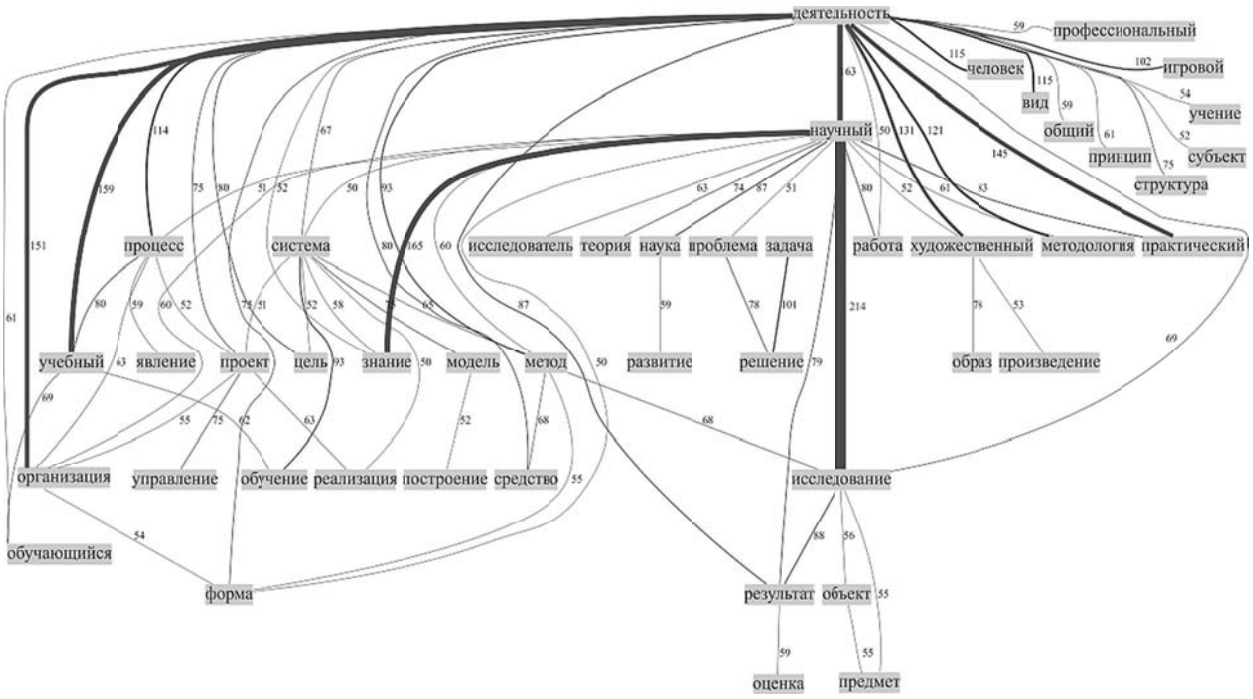

Рисунок 8 – Граф терминов ОМ

4.2.3 Анализ структурных центральностей терминов

Построим для ОМ аналогичные таблице 1 и рисунку 5 таблицу 2 (топ-10 терминов ОМ по каждому из показателей) и рисунок 9 (коэффициенты корреляции для показателей для ОМ). Имея сеть терминов, можно для каждого термина рассчитать его структурную значимость. Для оценки структурной значимости терминов используются следующие показатели: центральность по степени, центральность по PageRank, центральность по близости, центральность по посредничеству, а также влиятельность по акциональной модели.

Таблица 2 – Топ-10 терминов ОМ по каждому из показателей

№	Частота	Степень	Близость	Посредничество	Pagerank	Влиятельность
1	деятельность	деятельность	деятельность	деятельность	деятельность	деятельность
2	научный	процесс	научный	научный	научный	научный
3	система	цель	учебный	система	система	система
4	метод	форма	исследование	исследование	процесс	процесс
5	исследование	система	организация	учебный	исследование	исследование
6	процесс	результат	практический	художественный	метод	метод
7	проект	вид	художественный	знание	знание	знание
8	результат	общий	процесс	проект	результат	результат
9	знание	метод	знание	метод	учебный	форма
10	игра	научный	методология	объект	цель	учебный

На рисунке 9 представлены значения коэффициентов корреляции **С**пирмена для значений показателей терминов ОМ.

Как и для МКД, для случая ОМ высокие коэффициенты ранговой корреляции Спирмена (> 0,8) имеют место между частотой и такими переменными как степень, близость, PageRank и влиятельность. Поэтому можно подтвердить сделанный выше вывод, что для оценки структурной значимости терминов достаточно использовать частоту встречаемости терминов и центральность по посредничеству.

4.2.4 Выделение кластеров терминов ОМ в сети

Для выделения кластеров терминов в взвешенном графе использованы методы выявления сообществ теории социально-сетевого анализа. Наилучшие результаты для ОМ, как и для МКД, показал метод спинового стекла. При помощи метода спинового стекла получено четыре кластера терминов ОМ:

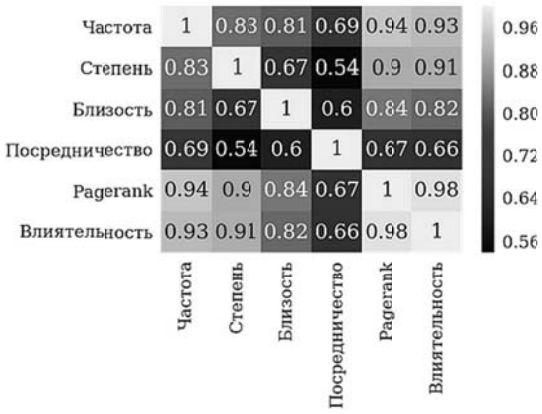

Рисунок 9 – Коэффициенты корреляций Спирмена для ОМ

1) деятельность, процесс, учебный, обучение, методология, организация, форма, средство, вид, принцип, практический, структура, образование, обучающийся, культура, содержание, подход, тип, профессиональный, учение, игровой, логический, школа, классификация, организационный, основание, способ, современный, логика, человеческий, характеристика, образовательный, учитель, особенность, педагог, книга;

2) научный, метод, исследование, результат, знание, наука, работа, теория, объект, развитие, понятие, предмет, явление, исследователь, опыт, теоретический, область, различный, основа, познание, эмпирический, практика, гипотеза, эксперимент, связь, учёный, изучение, предметный, свойство, измерение, закон, существенный, представление, признак, исследовательский, факт, отрасль, коллектив, идея;

3) система, проект, цель, проблема, модель, задача, решение, общий, оценка, условие, управление, критерий, анализ, этап, построение, определение, действие, конкретный, фаза, моделирование, реализация, выбор, стадия, проектирование, рефлексия, технология, шкала, информация, множество, программа, технологический, математический, механизм, элемент, функция, вариант, создание, сложный, план, состояние, операция, участник;

4) игра, человек, художественный, образ, отношение, уровень, определённый, искусство, правило, язык, смысл, субъект, группа, ситуация, эстетический, произведение, художник, ребёнок, качество, общество, жизнь, общественный, личность, класс, умение, труд, реальный, целое, индивидуальный, характер, коллективный, способность, познавательный.

Такие кластеры имеют хорошую содержательную интерпретацию и соответствуют структуре ОМ: кластер 1 – термины, характерные, в основном, для методологии практической и образовательной деятельности; кластер 2 – термины, характерные, в основном, для методологии научной деятельности; кластер 3 – «общеметодологические» термины; кластер 4 – термины, характерные, в основном, для методологии игровой и художественной деятельности.

4.3 Сравнение терминологических структур МКД и ОМ

Проведем сравнение между терминологическими структурами МКД и ОМ (см. методы сравнительного анализа в разделе 2), ограничимся при этом первыми 134 топ-терминами.

Рассмотрим распределение терминов по убыванию относительной частоты встречаемости в МКД и ОМ (частота встречаемости термина в соответствующем тексте нормируется на общее число слов в этом тексте, см. рисунок 10).

Как видно из рисунка 10, кривая кумулятивной частоты для МКД круче соответствующей кривой для ОМ и переходит в «плато» быстрее её. Это можно объяснить тем, что ПрО МКД более чёткая, в то время как в рамках ОМ рассматриваются разные виды деятельности, существенно отличающиеся друг от друга.

Будут ли часто встречающиеся в МКД термины столь же популярными в ОМ, и наоборот? Для этого упорядочим термины по убыванию частоты встречаемости и введём функцию сходства двух ранжирований S (по частоте встречаемости в МКД) и T (по частоте встречаемости в ОМ):

$$Sim(S,T,k) = \frac{|S_{1:k} \cap T_{1:k}|}{|S_{1:k} \cup T_{1:k}|}$$

где k – число топ-терминов, $S_{1:k}$ – множество первых k элементов ранжирования S, $T_{1:k}$ – множество первых k элементов ранжирования T. Зависимость сходства двух ранжирований от величины топа приведена на рисунке 11.

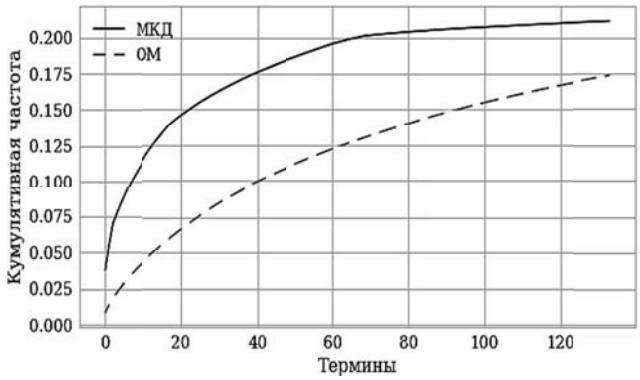

Рисунок 10 – Распределение терминов МКД и ОМ

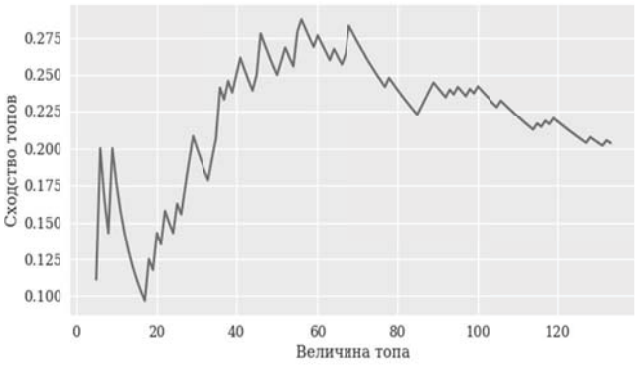

Рисунок 11 – Сходство топов МКД и ОМ

Из рисунка 11 видно, что при малых значениях параметра k множества часто встречающихся терминов МКД и ОМ сильно различаются (таким образом, наиболее популярные термины не являются одинаковыми в МКД и ОМ), затем достигается максимум сходства множеств (в этот момент в этих множествах появляются «популярные» и общие и для МКД, и для ОМ термины), а далее эти множества начинают все больше различаться (в них появляются узкоспециализированные термины).

В целом соотношения между множествами терминов МКД и ОМ показаны в таблице 3 (в скобках для каждого термина указана встречаемость в соответствующем тексте).

Таблица 3 – Ранжированные списки терминов

МКД\ОМ	ОМ\МКД	ОМ ∩ МКД
комплексный (1253)	научный (905)	деятельность (3148, 1440)
ресурс (421)	исследование (619)	элемент (1464, 95)
неопределённость (374)	проект (520)	система (503, 724)
цикл (318)	игра (470)	модель (558, 351)
жизненный (305)	учебный (443)	результат (399, 495)
операция (290)	человек (423)	метод (74, 652)
событие (195)	наука (419)	технология (562, 119)
элементарный (184)	обучение (404)	цель (259, 406)
спрос (168)	художественный (377)	процесс (135, 528)
причинно-следственный (164)	средство (325)	субъект (492, 155)
потребность (152)	решение (307)	управление (373, 242)
пул (119)	общий (296)	организация (271, 340)
анализируемый (102)	развитие (289)	структура (370, 232)
процессный (98)	образ (288)	знание (87, 472)
требование (94)	оценка (270)	работа (132, 406)
оптимизация (90)	практический (251)	предмет (279, 216)
эффективность (86)	понятие (226)	методология (153, 340)
регулярный (79)	уровень (214)	теория (114, 367)
ход (77)	определённый (211)	вид (156, 312)
фактор (77)	образование (210)	проблема (86, 372)
часть (76)	критерий (210)	объект (115, 332)
systems (75)	обучающийся (208)	задача (113, 321)
синтез (62)	культура (203)	форма (98, 335)
целеполагание (60)	явление (198)	условие (80, 249)
состояние (60)	искусство (196)	анализ (122, 206)
поведение (58)	определение (190)	отношение (78, 233)
результативность (54)	построение (190)	действие (124, 185)

МКД\ОМ	ОМ\МКД	ОМ ∩ МКД
сложность (46)	содержание (190)	принцип (47, 252)
управляющий (36)	опыт (176)	логический (155, 138)
эффект (22)	подход (174)	этап (81, 192)
агент (22)	конкретный (171)	правило (78, 185)
процессуальный (21)	область (170)	связь (129, 115)
эксплуатация (20)	различный (165)	фаза (89, 153)
параметр (20)	основа (164)	стадия (112, 127)
согласованность (19)	язык (161)	тип (74, 162)
планирование (19)	смысл (160)	характеристика (137, 96)
полезность (19)	группа (152)	основание (86, 127)
причина (18)	познание (151)	организационный (81, 128)
иерархический (18)	профессиональный (150)	проектирование (82, 126)
расширенный (16)	моделирование (150)	исследователь (10, 185)
признак (16)	реализация (150)	сложный (106, 81)
управляемый (16)	учение (147)	теоретический (12, 171)
complex (15)	ситуация (145)	рефлексия (48, 120)
архитектурный (15)	эстетический (143)	механизм (21, 97)
системотехника (15)	игровой (143)	логика (12, 105)
системный (15)	выбор (138)	
стандарт (15)	эмпирический (137)	
целенаправленный (14)	школа (137)	
программный (14)	произведение (136)	
прогноз (13)	практика (133)	
комплексирование (13)	классификация (132)	
операционный (13)	художник (132)	
ценность (13)	ребёнок (124)	
устранение (13)	способ (122)	
график (13)	качество (120)	
универсальный (12)	гипотеза (119)	
организованная (12)	современный (118)	
фрактальность (12)	эксперимент (118)	
смежный (12)	шкала (114)	
эскалирование (12)	общество (111)	
мера (12)	учёный (110)	
полный (12)	информация (110)	
эмерджентность (12)	жизнь (110)	
взаимосвязь (12)	множество (109)	
календарно-сетевой (11)	программа (107)	
operations (11)	человеческий (105)	
повторение (11)	математический (104)	
блюдо (11)	технологический (104)	
шаблон (11)	личность (103)	
стоимость (11)	изучение (103)	
business (11)	общественный (103)	
компетенция (11)	предметный (100)	
факт (11)	свойство (100)	
architecture (11)	измерение (97)	
увольнение (11)	класс (95)	
участник (11)	умение (93)	
конструкция (11)	функция (92)	
соотношение (10)	закон (92)	
составляющий (10)	труд (89)	
попытка (10)	вариант (89)	
technology (10)	образовательный (88)	
желаемый (10)	реальный (87)	
автоматический (10)	целое (86)	
прогнозирование (10)	индивидуальный (86)	
постановка (10)	создание (84)	
составной (10)	существенный (83)	
устройство (10)	характер (83)	
целостный (9)	учитель (80)	
методологический (7)	план (80)	

Из таблицы 3 видно (см. также рисунок 12), что общими для двух ПрО являются общие категории, в то время как для МКД специфичны «комплексность», «неопределённость», «технологии» и тесно связанные с ними термины; а для ОМ специфичны термины, отража-

ющие особенности различных видов деятельности (научной, образовательной и др.). На рисунке 12 приводится граф терминов, являющийся объединением графа терминов МКД (вес рёбер в таком графе ≥ 50) и графа терминов ОМ (вес рёбер в таком графе ≥ 50). Метки вершин включают в себя термин, частоту термина в МКД и частоту термина в ОМ. Красным цветом помечены вершины, относящиеся только к МКД, зелёным – только к ОМ, цвет остальных вершин «смешивается» из красного и зелёного в зависимости от соотношения частот терминов в МКД и ОМ. Толщина ребра графа зависит от суммы весов соответствующих рёбер в исходных графах, цвет рёбер, также как и цвет вершин, определяется соотношением этих весов.

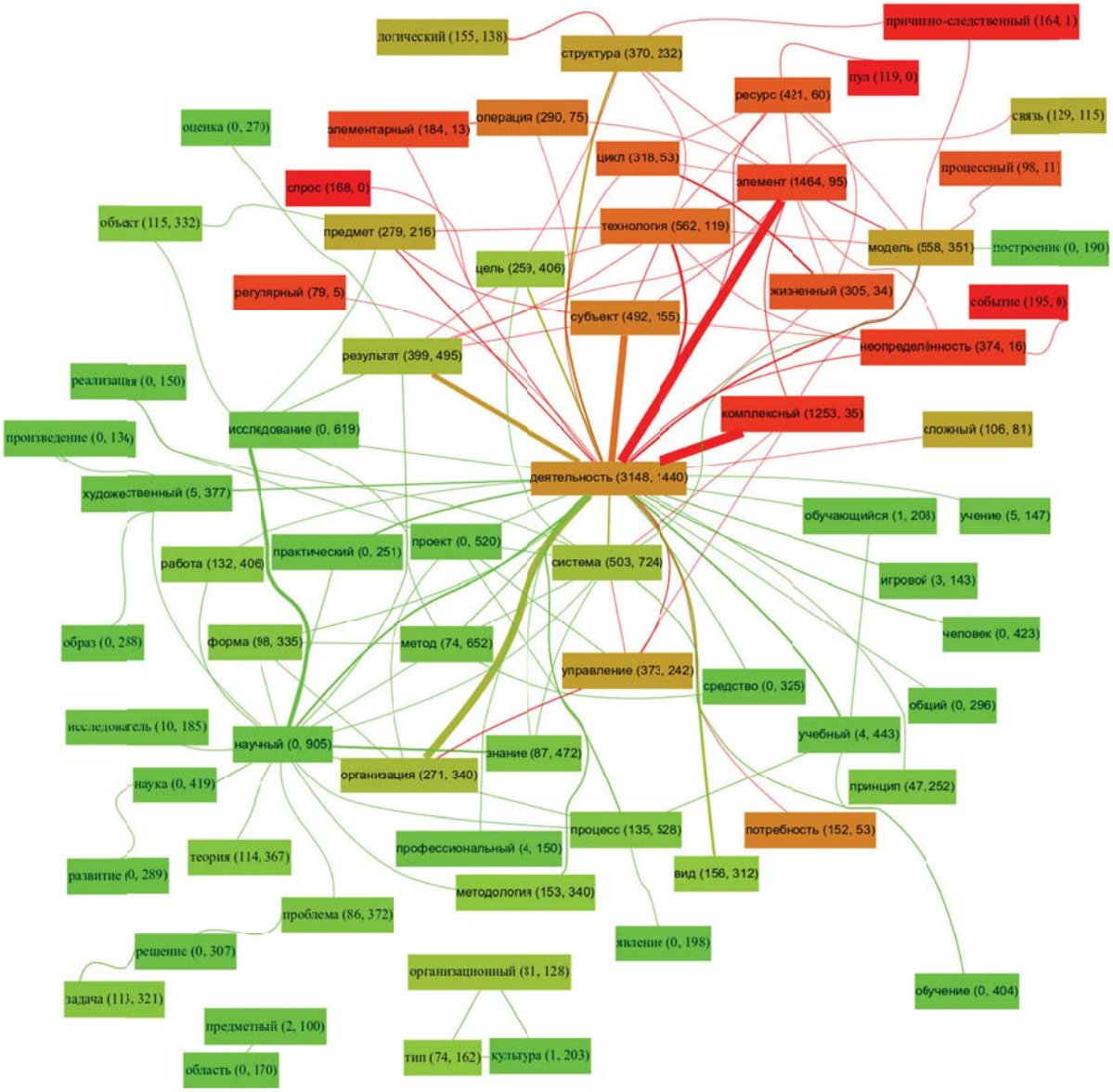

Рисунок 12 – Граф терминов МКД и ОМ

Заключение

Предложенные и апробированные на примере таких ПрО как МКД и ОМ методы извлечения и анализа терминологических структур позволяют:

- интерактивно (во взаимодействии с экспертом ПрО) извлекать термины ПрО из релевантной ей монографии (без дополнительного корпуса текстов);
- извлекать связи между терминами и формировать терминологическую структуру ПрО;
- оценивать структурную значимость терминов в терминологической структуре и определять наиболее подходящие показатели структурной значимости;
- выявлять содержательно интерпретируемые кластеры терминов терминологической структуры;
- проводить сравнительный анализ терминологических структур различных ПрО.

Список источников

[1] ***Новиков, А.М.*** Методология. 2-е изд. / А.М. Новиков, Д.А. Новиков – М.: УРСС, 2013. – 663 с.

[2] ***Губанов, Д.А.*** Методы анализа терминологической структуры предметной области (на примере методологии) / Д.А. Губанов, А.В. Макаренко, Д.А. Новиков // Управление большими системами. - 2013. - № 43. - С.5–33. - DOI: 10.1134/S0005117914120133.

[3] ***Белов М.В., Новиков Д.А***. Методология комплексной деятельности. – М.: Ленанд, 2018. – 320 с.

[4] ***Белов, М.В.*** Структура методологии комплексной деятельности / М.В. Белов, Д.А. Новиков // Онтология проектирования. – 2017. – Т.7, №4(26). - С.366-387. – DOI: 10.18287/2223-9537-2017-7-4-366-387.

[5] ***Astrakhantsev, N.A.*** Methods for Automatic Term Recognition in Domain-Specific Text Collections: A Survey / N.A. Astrakhantsev, D.G. Fedorenko, D.Yu. Turdakov // Programming and Computer Software, 2015. - Vol. 41. - No. 6. – P. 336–349. - DOI: 10.1134/S036176881506002X.

[6] ***Verberne, S.*** Evaluation and analysis of term scoring methods for term extraction / S. Verberne, M. Sappelli, D. Hiemstra et al**.** // Information Retrieval Journal, 2016. - Vol. 19. - Issue 5. - P.510-545. - DOI: 10.1007/s10791-016-9286-2.

[7] ***Астраханцев, Н.А.*** Автоматическое извлечение терминов из коллекции текстов предметной области с помощью Википедии / Н.А. Астраханцев // Труды ИСП РАН. - 2014. - Том 26. - Выпуск 4. - С.7–20. - DOI: 10.15514/ISPRAS-2014-26(4)-1.

[8] ***Manning, C.*** Foundations of Statistical Natural Language Processing / C. Manning, H. Schütze. – MIT press, 1999. - 620 p.

[9] ***Frantzi, K.*** Automatic recognition of multi-word terms: The C-value/NC-value method / K. Frantzi, S. Ananiadou, H. Mima // International Journal on Digital Libraries. - 2000. - Vol.3. – Issue 2. - P.115–130. - DOI: 10.1007/s007999900023.

[10] ***Браславский, П.И.*** Сравнение четырех методов автоматического извлечения двухсловных терминов из текста / П.И. Браславский, Е.А**.** Соколов // Компьютерная лингвистика и интеллектуальные технологии. Сборник трудов международной конференции Диалог 2006. Москва, 2006. - С.88–94.

[11] ***Justeson, J.S.*** Technical terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering / J.S. Justeson, S.M. Katz. - 1995. Vol.1. Issue 1. - P.9–27. - DOI: 10.1017/S1351324900000048.

[12] ***Hearst, M.A.*** Automatic acquisition of hyponyms from large text corpora // Proceedings of the 14th conference on Computational linguistics. - Vol. 2. – Association for Computational Linguistics, 1992. – P.539-545. DOI: 10.3115/992133.992154.

[13] ***Mikolov T. et al.*** Distributed Representations of Words and Phrases and their Compositionality // Advances in Neural Information Processing Systems. 2013. – P.3111-3119.

[14] ***Pelevina, M.*** Making Sense of Word Embeddings / M. Pelevina, N. Arefyev, C. Biemann, A. Panchenko // Proceedings of the 1st Workshop on Representation Learning for NLP. Berlin, Germany, August 11th, 2016. - P.174–183. - DOI: 10.18653/v1/W16-1620.

[15] ***Губанов, Д.А.*** Акциональная модель влиятельности пользователей социальной сети / Д.А. Губанов, А.Г. Чхартишвили // Проблемы управления. – 2014. - № 4. С.20–25. - DOI: 10.1134/S0005117915070139.

[16] ***Wasserman, S.*** Social network analysis: Methods and applications / S. Wasserman, K. Faust. – Cambridge university press, 1994. - 857 p. - DOI: 10.1017/CBO9780511815478.

[17] ***Kempe, D.*** Maximizing the Spread of Influence through a Social Network / D. Kempe, J. Kleinberg, E. Tardos // Theory of Computing. - 2015. Vol. 11. No. 4. - P.105–147. - DOI: 10.4086/toc.2015.v011a004.

[18] ***Wermter, J.*** You can't beat frequency (unless you use linguistic knowledge): a qualitative evaluation of association measures for collocation and term extraction / J. Wermter, U. Hahn // Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics / Association for Computational Linguistics. - 2006. - P.785-792. - DOI: 10.3115/1220175.1220274.

[19] ***Reichardt, J.*** Statistical Mechanics of Community Detection / J. Reichardt, S. Bornholdt // Physical Review E. – 2006. – V. 74. – No. 1. – P.16110. - DOI: 10.1103/PhysRevE.74.016110.

EXTRACTION AND ANALYSIS METHODS FOR THE TERMINOLOGICAL STRUCTURES OF RELATED SUBJECT AREAS

D.A. Gubanov[1], D.A. Novikov[2]

Trapeznikov Institute of Control Sciences Russian Academy of Sciences, Moscow, Russia
[1]dmitry.a.g@gmail.com, [2]novikov@ipu.ru

Abstract

We suggest an automated expert approach to extract terminological structures from monographs. It is postulated that the content of scientific subject area is fully reflected in some monograph, it allows to extract and analyze terminological structures without using a pre-prepared corpus of specialized domain documents. Within the framework of the approach, we consider the procedure of iterative extraction and analysis of terminological structures that regulates the work of an expert with a monograph and allows him to control the process. Comparative analysis methods for the terminology structures of two related subject areas are proposed. The possibilities of the developed approach are demonstrated on the example of the general methodology and methodology of complex activities.

Key words: *subject area, terminology structure, graph theory, general methodology, methodology of complex activity.*

Citation: *Gubanov DA, Novikov DA.* Extraction and analysis methods for the terminological structures of related subject areas [In Russian]. *Ontology of designing*. 2018; 8(3): 347-365. - DOI: 10.18287/2223-9537-2018-8-3-347-365.

References

[1] ***Novikov AM, Novikov DA***. Methodology [In Russian]. – Moscow.: URSS, 2013. – 663 p.
[2] ***Gubanov DA, Makarenko AV, Novikov DA***. Analysis methods for the terminological structure of a subject area // Automation and Remote Control. 2014; 75(12): 2231-2247. DOI: 10.1134/S0005117914120133.
[3] ***Belov MV, Novikov DA.*** Methodology of complex activity [In Russian]. – Moscow: Lenand; 2018.
[4] ***Belov MV, Novikov DA***. Structure of methodology of complex activity [In Russian]. Ontology of designing. 2017; 7(4): 366-387. DOI: 10.18287/2223-9537-2017-7-4-366-387.
[5] ***Astrakhantsev NA, Fedorenko DG, Turdakov DYu***. Methods for Automatic Term Recognition in Domain-Specific Text Collections: A Survey // Programming and Computer Software, 2015; 41(6): 336–349. DOI: 10.1134/S036176881506002X.
[6] ***Verberne S., Sappelli M, Hiemstra D. et al***. Evaluation and analysis of term scoring methods for term extraction // Information Retrieval Journal, 2016; 19(5): 510-545. DOI: 10.1007/s10791-016-9286-2.
[7] ***Astrakhantsev NA.*** Automatic term recognition in a domain-specific text collection using Wikipedia [In Russian]. Proceedings of ISP RAS, 2014; 26(4): 7–20. - DOI: 10.15514/ISPRAS-2014-26(4)-1.
[8] ***Manning C, Schütze H***. Foundations of Statistical Natural Language Processing. – MIT press, 1999. – 620 p.
[9] ***Frantzi K, Ananiadou S, Mima H***. Automatic recognition of multi-word terms: The C-value/NC-value method // International Journal on Digital Libraries. – 2000; 3(2): 115–130. DOI: 10.1007/s007999900023.
[10] ***Braslavskii PI, Sokolov EA.*** Comparison of four methods for automatic recognition of two-word terms in text [In Russian]. Computational Linguistics and Intellectual Technologies, 2006. - P.88–94.
[11] ***Justeson JS, Katz SM***. Technical terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering. 1995; 1(1): 9–27. DOI: 10.1017/S1351324900000048.
[12] ***Hearst MA.*** Automatic acquisition of hyponyms from large text corpora // Proceedings of the 14th conference on Computational linguistics. - Vol. 2. – Association for Computational Linguistics, 1992. – P.539-545. DOI: 10.3115/992133.992154.
[13] ***Mikolov T.*** **et al.** Distributed Representations of Words and Phrases and their Compositionality // Advances in Neural Information Processing Systems. 2013. – P.3111-3119.

[14] ***Pelevina M*, *Arefyev N*, *Biemann C*, *Panchenko A*.** Making Sense of Word Embeddings / Proceedings of the 1st Workshop on Representation Learning for NLP. Berlin, Germany, August 11th, 2016. P. 174–183. DOI: 10.18653/v1/W16-1620.

[15] ***Gubanov DA*, *Chkhartishvili AG***. An actional model of user influence levels in a social network // Automation and Remote Control. 2015; 76(7): 1282–1290. DOI: 10.1134/S0005117915070139.

[16] ***Wasserman S*, *Faust K***. Social network analysis: Methods and applications. – Cambridge university press, 1994. – 857 p. DOI: 10.1017/CBO9780511815478.

[17] ***Kempe D*, *Kleinberg J*, *Tardos E*.** Maximizing the Spread of Influence through a Social Network // Theory of Computing. 2015; 11(4): 105–147. DOI: 10.4086/toc.2015.v011a004.

[18] ***Wermter J*, *Hahn U***. You can't beat frequency (unless you use linguistic knowledge): a qualitative evaluation of association measures for collocation and term extraction // Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics / Association for Computational Linguistics. 2006. - P.785-792. DOI: 10.3115/1220175.1220274.

[19] ***Reichardt J*, *Bornholdt S*.** Statistical Mechanics of Community Detection // Physical Review E. 2006; 74(1): 16110. DOI: 10.1103/PhysRevE.74.016110.

Сведения об авторах

Губанов Дмитрий Алексеевич, 1984 г. рождения, кандидат технических наук, старший научный сотрудник Института проблем управления Российской академии наук, автор более 60 научных работ. Область научных интересов – информационное управление в социальных сетях, социально-сетевой анализ, теория управления социально-экономическими системами.

Dmitry Alexeyevich Gubanov was born in 1984. He received Candidate of Science degree in engineering (Ph.D. degree) from the Institute of Control Sciences of the Russian Academy of Sciences, Moscow, in 2009. At present, he is a Senior Researcher with the Institute of Control Sciences of the Russian Academy of Sciences, Moscow, Russia. He is the author of more than 60 publications. His research interests include social network analysis, big data analytics, decision-making and mechanisms of control of social and economic systems.

Новиков Дмитрий Александрович, 1970 г. рождения, доктор технических наук, профессор, член-корреспондент РАН, директор Института проблем управления Российской академии наук, заведующий кафедрой интегрированных киберсистем Московского физико-технического института. Автор более 500 научных работ по теории управления системами междисциплинарной природы, в том числе – по системному анализу, теории игр, принятию решений, управлению проектами и математическим моделям механизмов управления социально-экономическими системами.

Dmitry Aleksandrovich Novikov was born in 1970. He is Doctor of Science in engineering. At present, he is Director of the Institute of Control Sciences of the Russian Academy of Sciences, professor, Corresponding Member of Russian Academy of Sciences. He is the author of more than 500 publications in the field of control of systems of interdisciplinary nature, including system analysis, game theory, project management, decision-making and mechanisms of control of social and economic systems.