ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 31 за 2017 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

**Орлов Ю.Н., Парфенова Ю.А.**

# Анализ структуры онтологического графа толкового словаря

***Рекомендуемая форма библиографической ссылки:*** Орлов Ю.Н., Парфенова Ю.А. Анализ структуры онтологического графа толкового словаря // Препринты ИПМ им. М.В.Келдыша. 2017. № 31. 25 с. doi:10.20948/prepr-2017-31
URL: http://library.keldysh.ru/preprint.asp?id=2017-31

**О р д е н а  Л е н и н а**
**ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ**
**имени М.В.Келдыша**
**Р о с с и й с к о й  а к а д е м и и  н а у к**

**Ю.Н. Орлов, Ю.А. Парфенова**

# Анализ структуры онтологического графа толкового словаря

**Москва — 2017**

**Орлов Ю.Н., Парфенова Ю.А.**

Анализ структуры онтологического графа толкового словаря

В работе описывается алгоритм и структура программы построения больших онтологий на примере русского толкового словаря. Вводятся понятия, отражающие специфику ориентированных графов с большим числом циклов, вершин и ребер. Определяются: иерархия вершины по отношению к корневой вершине (порождающему слову), бассейн слова, оболочка бассейна, базис бассейна и набор слов, порождающих бассейн. Исследуются статистические свойства графа онтологии бассейна для одного из слов ряда общих классификаторов (в данной работе это слово «предмет»). Анализируется также зависимость числа вершин, связей и глубины онтологии от уровня иерархии слова в бассейне.

***Ключевые слова:*** онтологический граф, статистика циклов, иерархия вершины, бассейн слова, базис бассейна

**Orlov Yu.N., Parfenova Yu.A.**

Ontological graph structure analysis for dictionary

The paper describes the algorithm and structure of the program for construction of large ontologies, for example, Russian dictionary. The concepts of specificity of oriented graphs with a large number of loops, vertices and edges are introduced. The vertex hierarchy, word basin, basin shell and basin volume are determined. The statistical properties of ontological graph are investigated.

***Key words:*** ontological graph, cycle statistics, vertex hierarchy, word basin, basin basis

Работа выполнена при поддержке гранта РНФ, проект № 14-21-00025

**Содержание**

Введение ..... 3
1. Основные задачи анализа структуры связей толкового словаря ..... 6
2. Теоретические оценки параметров графа словаря ..... 8
3. Алгоритмы программы статистического анализа графа онтологии ..... 10
4. Построение онтологического графа слова «предмет» ..... 12
5. Оболочка и граница бассейна ..... 19
6. Модель базиса ядра словаря ..... 20
7. Результаты анализа и дальнейшие перспективы ..... 22
Литература ..... 25

**Введение**

Цель настоящего исследования состоит в создании программной среды для обработки и анализа направленно связанной системы большого количества объектов, так называемого онтологического графа [1]. Согласно толковому словарю Ожегова и Шведовой [2], «*онтология – это философское учение об общих категориях и закономерностях бытия, существующее в единстве с теорией познания и логикой*». Мы специально привели это определение, поскольку в дальнейшем предполагаем изучить онтологические связи, формирующиеся внутри конкретно этого толкового словаря. Кроме чисто формальной задачи анализа и структурирования больших данных, естественно возникающей при описании связей между объектами, для которых введено отношение смежности, есть и задача определения близости двух или более систем объектов в случае, когда интересующей метрикой является мера их семантической однородности, понимаемая в некотором специальном смысле. Например, в [3] приводится следующее определение: «*онтология – это подробная спецификация структуры определенной проблемной области*». Возникает вопрос: об одном и том же говорится в двух приведенных определениях? Ведь в них нет ни одного совпадающего слова. В другой статье [4] читаем: «*онтология – это частная система категорий, относящаяся к некоторой области знаний*». За исключением слова «область», это определение по словарному составу также полностью отличается от предыдущих двух. Следует выяснить, насколько изменяется смысл высказывания, если его переформулировать применительно к конкретной области исследования или, напротив, провести некое его обобщение. И возможно ли вообще ввести метрику, позволяющую установить смысловую близость подобных высказываний без обращения к мнению эксперта?

Существует множество подходов к определению семантической близости, и мы не имеем цели детально их анализировать. Обзор основных методов приведен в работе П.Е. Велихова [5]. Существенно здесь то, что каждый из подходов становится малоэффективным при увеличении длины анализируемого текста. Снижение эффективности проявляется в том, что разные тексты начинают идентифицироваться как одинаковые по смыслу и наоборот.

Одни исследователи предлагают определить «смысл» высказывания или совокупности высказываний путем экспертного вычленения тех или иных категориальных кластеров, другие – через выделение ключевых слов и сопоставление им некоторых векторов, по углу между которыми можно судить о близости текстов, третьи – через гипотезу о случайном блуждании вершин объясняющих слов на графе семантических связей и введении так называемой меры Ньюмана [6]. Однако «разность» и «одинаковость» смыслов в таких подходах зависят от экспертного мнения читателя: один посчитает, что, например, некоторые два детективных романа Фрэнсиса Дика одинаковые,

потому что оба «о лошадях», а другой посчитает их разными, потому что в одном лошадь была гнедая, а в другом – белая в яблоках.

Следовательно, необходимо опереться на некоторый условный эталон смысла понятия или высказывания, который, строго говоря, определяется в соответствии со статьями толкового словаря, и потому должен сводиться к сочетанию некоторых базовых взаимосвязанных понятий. Эти понятия в предположении существования объективного смысла должны регулярным образом выделяться из словаря без апелляции к мнению экспертов.

Таким образом, для цели сравнения высказываний может оказаться полезным создание определенного базиса понятий, по которому в конечном итоге будут разложены составляющие высказывание слова, если между словами задано отношение смежности. Возможно, что интерес будут представлять определенные кластеры понятий, выделенные по принципу цикличности внутренних связей. Близость между высказываниями тогда может быть оценена по степени пересечения объектов, каждый из которых получен объединением кластеров, отвечающих входящим в высказывание понятиям.

Концептуально наша работа близка к статье [1], в которой дан теоретический анализ методологии построения моделей компьютерной онтологии. В указанной работе, в частности, считается, что для установления связей между словами следует использовать набор доступных толковых словарей, статьи которых необходимо преобразовать определенным образом, чтобы корректно проводить их компьютерную обработку. Однако непосредственно анализ словаря в [1] не содержится.

Надо сказать, что на данный момент существует множество работ, в которых даются рекомендации и приводятся примеры моделей для реализации онтологий словарных систем. Например, в работах [7] и [8] описана методология построения системы машинного анализа структуры словаря и принципов построения соответствующего графа, приводятся алгоритмы процесса редактирования словарных статей, процедур анализа, а также возможные применения такой системы. Однако результаты применения подобных программ к описанию именно толковых словарей на практике до сих пор отсутствуют. Возможно, дело здесь в том, что структура получаемого «словарного графа» крайне неудобна.

Традиционно онтологические графы применяются для моделирования социальных сетей, понимаемых в широком смысле: это и схемы управления промышленным предприятием, и схемы транспортных путей, и конструирование ссылок на статьи справочника, в частности, описание структуры связей интернет-ресурсов, а также схемы связей между адресатами разного рода сообщений.

Управленческие онтологии близки к идеальным, каковыми считаются графы без циклов. Пример такого графа приведен на рис. 1.

Структура относительно простой социальной сети (конспиративного управления агентами) менее идеальна (рис. 2), но и она имеет вполне четкие законы функционирования.

Транспортная сеть имеет обычно радиально-кольцевую структуру с узлами пересадок, т.е. вполне четко структурирована. Часто ребра транспортного графа берутся с весами, характеризующими фактическое расстояние между пунктами назначения. Регламент функционирования такой сети (т.е. граф управления) представляет собой простую онтологию по типу рис. 1.

Социальная сеть адресатов представляет собой объединение кластеров по интересам, причем внутри кластера граф близок к полному (все вершины связаны одна с другой).

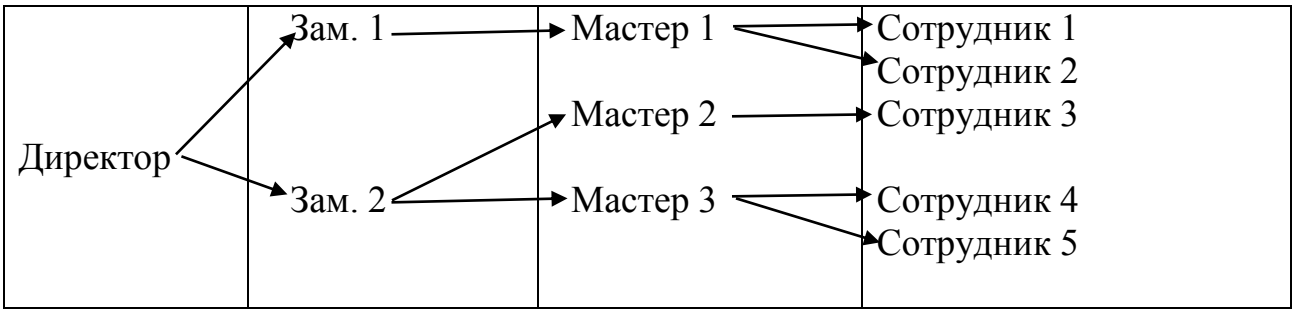



Рис. 1. Пример графа структуры управления производством

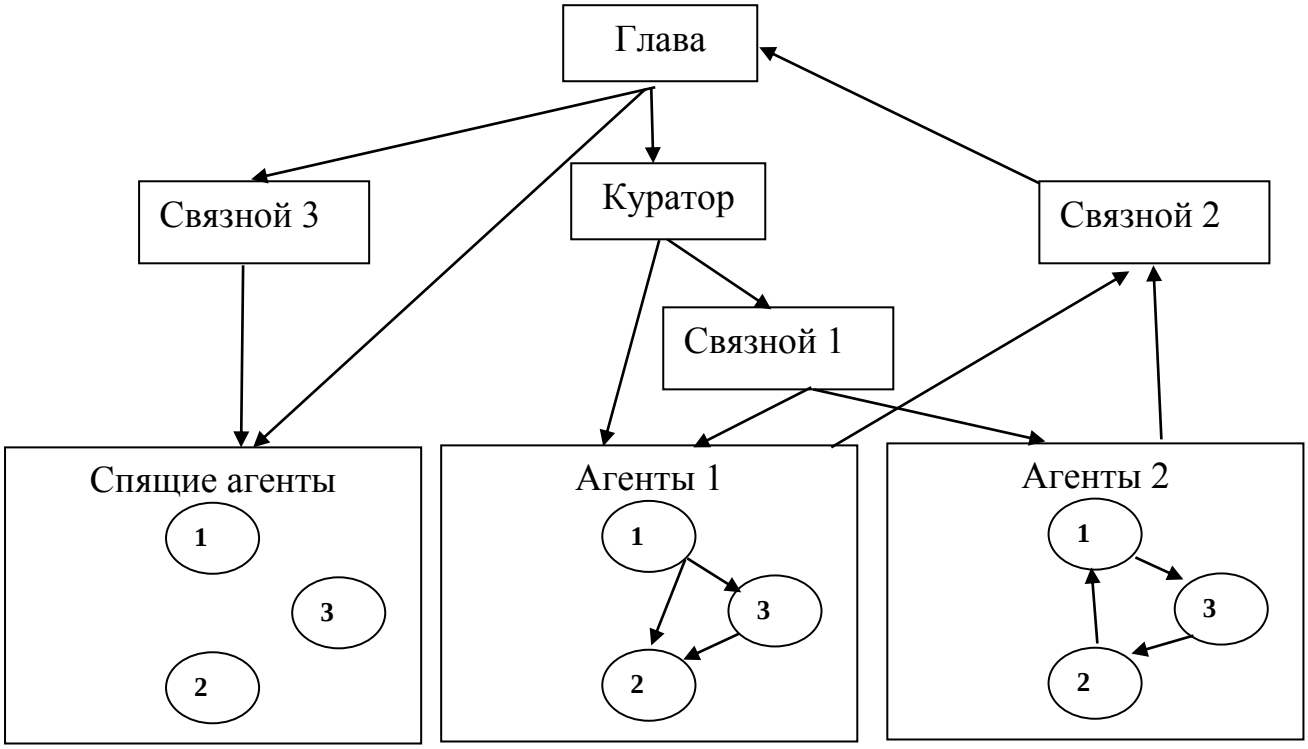



Рис. 2. Пример графа структуры простой социальной сети

Граф связей между словами в словаре принципиально отличается от вышеописанных структур. Он близок к случайному графу, хотя по факту и не является таковым. Его характерной чертой является то, что из каждой вершины выходит хотя бы одно ребро, однако существуют вершины, для которых нет ни одной входящей связи. Это означает, что каждый путь, исходящий из любой

вершины такого графа, замыкается в цикл, но, возможно, не обязательно в той вершине, из которой вышел.

Если занумеровать слова в словаре, например, в алфавитном порядке, то структурно схема словаря может быть представлена матрицей из нулей и единиц ($0_{jk}$ – вершины j, k не соединены, $1_{jk}$ – такая связь имеется) размером $N \times N$, где $N$ есть число слов в словаре. Каждое слово объясняется в словаре с помощью предложения или предложений, совокупно содержащих порядка 10 слов после соответствующей обработки. Следовательно, структурная матрица словаря весьма разрежена, поскольку для 20-30 тыс. слов (общеупотребительный лексикон) в каждой строке этой длины будет находиться в среднем 10 единиц, а остальные элементы – нули. Более того, поскольку существуют слова, к которым словарные статьи не обращаются за разъяснением, а лишь сами эти слова объясняются через другие слова, то соответствующий столбец структурной матрицы будет нулевым. При этом базисный минор структурной матрицы может не отвечать «семантической правде жизни», поскольку не все слова могут быть связаны одно с другим с образованием осмысленного высказывания. Следовательно, требуется проводить более тонкий статистический анализ связей словарной онтологии, для чего следует разработать как концепцию, так и реализующий ее численный алгоритм, эффективно работающий в среде с Большими Данными.

Наличие соответствующей программы формирования структуры со встроенными алгоритмами анализа графов позволит существенно упростить процесс анализа сложных сетей, поиска кластеров, определенных путей (цепей или циклов), выделения зависимости числа ссылок от уровня иерархии связи и т.д. В свою очередь, это даст возможность проводить более масштабные лингвистические, социальные и другие исследования, а также уточнить методы семантического анализа.

## 1. Основные задачи анализа структуры связей толкового словаря

Процесс объяснения смысла слова с помощью статей толкового словаря состоит в следующем. Стартуя с какого-либо слова (нулевой уровень иерархии объяснения применительно к данному слову), мы читаем объясняющую его словарную статью (первый уровень иерархии), после чего каждому слову в этой статье ищем объяснение в том же словаре (второй уровень иерархии) и т.д. Предполагается, естественно, что в словаре не используются слова, значения которых не объяснены. Процедура объяснения заканчивается, когда каждая ветвь заворачивается в цикл, т.е. когда на определенном уровне иерархии не возникает слов, на которые не было ссылок с предыдущих уровней.

При проведении описанной процедуры естественно возникают следующие понятия и связанные с ними вопросы.

Пусть $n(A)$ есть глубина иерархии слова $A$, при которой все пути замкнулись в циклы, и $N(A)$ - число вершин в полученном графе, который будем называть *бассейном слова* $A$ и обозначать $B(A)$. Само слово $A$ будем

называть *словом, порождающим бассейн* $B(A)$. При этом другие слова $A_i$, входящие в бассейн $B(A)$, также могут иметь свои бассейны $B(A_i)$, которые по построению содержатся в бассейне $B(A)$. Объединение $\bigcup_i B(A_i)$ содержится в бассейне $B(A)$. Если степень вершины слова $A$ по входящим ребрам больше нуля, то $B(A) = \bigcup_i B(A_i)$. Каждая вершина $A_i$ бассейна $B(A)$ характеризуется парой чисел: количеством $n$ выходящих ребер и количеством $m$ входящих. Вершину с указанием ее степени обозначаем $A_i(n,m)$. Число входящих ребер в данную вершину, как и число выходящих образуют целочисленные векторы $\mathbf{n} = (n_1,...,n_{n(A)})$ и $\mathbf{m} = (m_1,...,m_{n(A)})$, размерность которых равна в общем случае глубине иерархии $n(A)$.

Каково распределение входящих и выходящих ребер по уровням иерархии? Каково распределение вершин по числу выходящих и входящих ребер? Как меняется доля новых слов с увеличением уровня иерархии? Какова зависимость числа ребер от уровня иерархии? Эти вопросы тесно связаны с априорной оценкой средней ширины онтологии как отношения объема бассейна (абсолютная ширина) и числа уровней иерархии.

Дальнейший анализ предполагает сбор основных статистик, которыми характеризуется онтология бассейна $B(A)$. Это распределение входящих в него бассейнов, включая $B(A)$, по объемам $N(A_i)$, распределение бассейнов по глубинам иерархий $n(A_i)$, а также корреляционная функция объема бассейна и глубины иерархии.

Большой интерес представляет распределение бассейна $B(A_i)$ по числу различных содержащихся в нем циклов. Цикл, начинающийся (и заканчивающийся) в вершине $A_i' \in B(A_i)$, будем обозначать $C(A_i; A_i')$. Для однозначной идентификации цикла его начальной точкой будем считать вершину, ближайшую по уровню иерархии к порождающей вершине бассейна. Длиной $L(A_i; A_i')$ цикла $C(A_i; A_i')$ будем называть число ребер в цепи без самопересечений, первая вершина которой $A_i' \in B(A_i)$ является также концом цепи. Если рассматривается один определенный бассейн, то цикл длиной $n$ ребер, проходящий через заданные вершины, обозначаем без указание на слово, порождающее бассейн: $C_n(A_1, A_2,...,A_n)$. Вопросы, требующие изучения в рамках этой группы понятий, следующие. Насколько функция распределения $U_{A_i}(L)$ циклов бассейна $B(A_i)$ по длинам близка к функции распределения $U_A(L)$ бассейна, который включает в себя все бассейны $B(A_i)$? Существует ли корреляция между объемом бассейна, числом содержащихся в нем циклов и максимальной длиной цикла?

Поиск предположительного базиса бассейна, а также базиса всего словаря основывается на топологии подграфов данного бассейна. Рассмотрим все бассейны $B(A_i)$ внутри данного $B(A)$. Есть ли у них непустое пересечение?

Именно, что представляет собой множество $\Omega = \bigcap_{i=1} B(A_i)$? Насколько бассейн $B(\Omega)$ отличается от $B(A)$? Можно ли выделить базисные слова, которые не связаны между собой цепочкой ребер? Если существует разумное определение базиса словаря, можно ли тогда каждому слову сопоставить некоторую числовую функцию, характеризующую смысл этого слова? Исследованию некоторых из поставленных вопросов и посвящена данная работа.

## 2. Теоретические оценки параметров графа словаря

Оценим среднюю ширину графа онтологии для некоторого слова $A$ из словаря. Абсолютной шириной онтологии называется объем бассейна $N(A)$. Пусть на каждом $i$-м уровне иерархии содержится $k_i$ новых слов, возникших вследствие ссылок на них с предыдущего уровня. Слова, ссылки на которые уже существовали ранее, в новый уровень не поступают. Число уровней иерархии равно $i = 1,2,\ldots,n(A)$, причем $\sum_{i=1}^{n(A)} k_i = N(A) - 1.$

Пусть каждому слову отвечает в среднем $\nu$ слов следующей иерархии. Если бы не было ограничения на полное число слов в словаре и онтология была бы идеальной, на каждом следующем уровне иерархии число новых слов возрастало бы в $\nu$ раз, т.е. $k_{t+1} = \nu k_t$ при «начальном» условии $k_0 = 1$. Однако фактически не все слова следующего уровня иерархии являются новыми, так что частично начинают повторяться слова с предыдущих уровней. В результате идеальная зависимость $k_t = \nu^t$ нарушается. Отклонение фактической зависимости от вышеуказанной идеальной определяется долей уже использованных слов в полном объеме бассейна. На $t$-ом уровне иерархии эта доля равна

$$\mu(t) = \frac{1}{N(A)} \sum_{i=0}^{t} k_i \,. \qquad (1)$$

Будем предполагать, что количество вершин на данном уровне $t$ иерархии бассейна определяется формулой

$$k_{t+1} = \nu k_t V\big(\mu(t)\big), \qquad (2)$$

где $V(\mu)$ есть некоторая функция, монотонно убывающая от 1 для $\mu(0) = 1/N$ до 0 для $\mu(n) = 1$ на некоторой максимальной глубине иерархии $n$. Строго говоря, поскольку количество слов выражается целым неотрицательным числом, то в правой части (2) следует взять целую часть получаемого выражения. Средней шириной графа для данного слова $A$ называется среднее число вершин, приходящееся на один уровень иерархии:

$$s(A) = \frac{N(A)}{n(A)}. \qquad (3)$$

Величина $n(A)$, т.е. число уровней иерархии, определяется из условия положительности выражения (2). Оценим $n(A)$ из дифференциального уравнения, в которое переходит (2) в случае непрерывного индекса. Далее указание на конкретное слово $A$ в обозначениях этого раздела для краткости опускаем. Имеем

$$\begin{gathered} k(t+1)-k(t) \equiv \frac{dk(t)}{dt} = k(t)\cdot\left(\nu V\left(\mu(t)\right)-1\right), \\ \mu(t) = \frac{1}{N}\cdot\left(1+\int_0^t k(\tau)d\tau\right), \quad \int_0^n k(\tau)d\tau = N-1. \end{gathered} \tag{4}$$

Из (4) следует, что

$$\frac{d\mu(t)}{dt} = \frac{k(t)}{N}, \tag{5}$$

так что уравнение (4) для $dk/dt$ в терминах переменной $\mu(t)$ (доли накопленных слов бассейна) примет вид

$$\mu'' = \mu'\cdot\left(\nu V\left(\mu\right)-1\right). \tag{6}$$

Поскольку номер $t$ иерархии в уравнение (6) явно не входит, порядок уравнения можно понизить, вводя функцию $g(\mu)=\mu'$. Внутри бассейна, где мы и рассматриваем заселенность $t$-го уровня иерархии, $g(\mu)>0$, поскольку, согласно (1), $\mu(t)$ есть строго возрастающая функция. Тогда после подстановки в (6) соотношения $\mu''=gg'$ получаем

$$\frac{dg}{d\mu} = \nu V(\mu)-1, \tag{7}$$

и решение сводится к простой квадратуре. Максимальная заселенность уровня определяется условием $g'=0$. В терминах (7) это означает, что накопленная доля $\tilde{\mu}$ в точке максимума ширины графа определяется из уравнения

$$V\left(\tilde{\mu}\right) = \frac{1}{\nu}. \tag{8}$$

Выбор функции $V$, естественно, может заметно повлиять на численные оценки параметров графа. Например, если принять линейную аппроксимацию $V(\mu)=(1-\mu)N/(N-1)$, то для функции $d\mu/dt$ будет получаться логистическое уравнение, свойства которого хорошо изучены. Однако сравнение получаемых в результате такой модели оценок с практически наблюдаемыми свойствами графа показывает, что модель весьма не точна. Как мы увидим далее, одной из актуальных аппроксимаций функции $V$ является логарифмическая:

$$V(\mu) = -\frac{\ln\mu}{\ln N}. \tag{9}$$

Для нее получаем из (7) решение в виде

$$g(\mu) = \frac{d\mu}{dt} = C - \left(1-\frac{\nu}{\ln N}\right)\mu - \frac{\nu}{\ln N}\mu\ln\mu,$$

где постоянная интегрирования $C$ находится из условий $\mu(0)=1/N,\ \mu'(0)=1/N$. Последнее условие есть следствие (5) с учетом того, что $k(0)=1$. В результате получаем

$$\frac{d\mu}{dt}=\frac{1}{N}\left(2-\nu-\frac{\nu}{\ln N}\right)-\mu(1-\lambda)-\lambda\mu\ln\mu,\ \lambda=\frac{\nu}{\ln N}. \qquad (10)$$

Общего решения уравнения (10) в элементарных функциях получить не удается. Однако можно рассмотреть приближение, когда $N>>1$, что выполняется на практике. Тогда свободным членом в (10) можно пренебречь, после чего решение этого уравнения может быть найдено аналитически:

$$\mu(t)=\exp\left(\left(\frac{1-\nu}{\lambda}-1\right)e^{-\lambda t}-\frac{1}{\lambda}+1\right). \qquad (11)$$

Эта модель довольно хорошо соответствует реальному распределению слов по иерархическим уровням.

### 3. Алгоритмы программы статистического анализа графа онтологии

При обработке текста словаря, построении и анализе полученного онтологического графа использовались алгоритмы лемматизации и систематизации для анализа текстов [9-11], а также алгоритм Косараю для поиска сильносвязных компонент и алгоритм Флойда-Уоршелла [12-14] для поиска кратчайших путей в графе. Данные алгоритмы были реализованы на языке Python.

При помощи языка Javascript была создана среда для ручной обработки словарных статей. Она принимает на вход данные, полученные при помощи описанных выше алгоритмов автоматической обработки, а также оригинальный словарь существительных. Страница редактора представляет собой определение слова из словаря: оригинал и соответствующие полученные при машинной обработке словоформы, которые можно при желании отредактировать. При этом, для упрощения задачи, в программе есть функция подсказки: в процессе набора слова отображаются возможные варианты редактирования (распознанные ранее существительные), а результат редактирования обобщается на весь словарь, что в разы ускоряет процесс ручной обработки. В любой момент текущее состояние системы можно зафиксировать в виде файла в формате JSON.

Также, с помощью Javascript была создана программа для визуализации связей в графе определений. Для удобства восприятия изображения построение связей было реализовано следующим образом.

1. Исходный словарь был предобработан на языке Python в формат, удобный для получения по заданному слову его входящих и исходящих связей.
2. Полученные при предобработке данные подаются на вход программы, в которой в предложенную форму вводится слово, связи которого нас интересуют, а также интересующая нас глубина этих связей.

3. На экране отрисовывается интересующая нас часть онтологического графа в удобном формате, который также позволяет анализировать связи для каждого отображаемого на ней слова.

Время работы алгоритма для различных параметров системы и браузера может отличаться, но в среднем (Google Chrome, Win8, Intel i3) связи для глубины иерархии до 2 с числом стрелок до 3500 отрисовываются в пределах 1 секунды, для глубины 3 – 1-2 сек., для глубины 4-5 (4000-6500 стрелок) – 3-5 сек.

Обе эти программы оптимизированы для работы под Google Chrome 56 и выше, Windows 8/10. Благодаря использованию Javascript они являются кроссплатформенными и не требуют установки.

На рис. 3 приведен пример работы алгоритма по выявлению входящих и выходящих связей, ассоциированных с определенным словом словаря в соответствии с его словарными статьями. Ориентация графа отображается цветом: выходу ребра отвечает синий цвет, входу – красный.

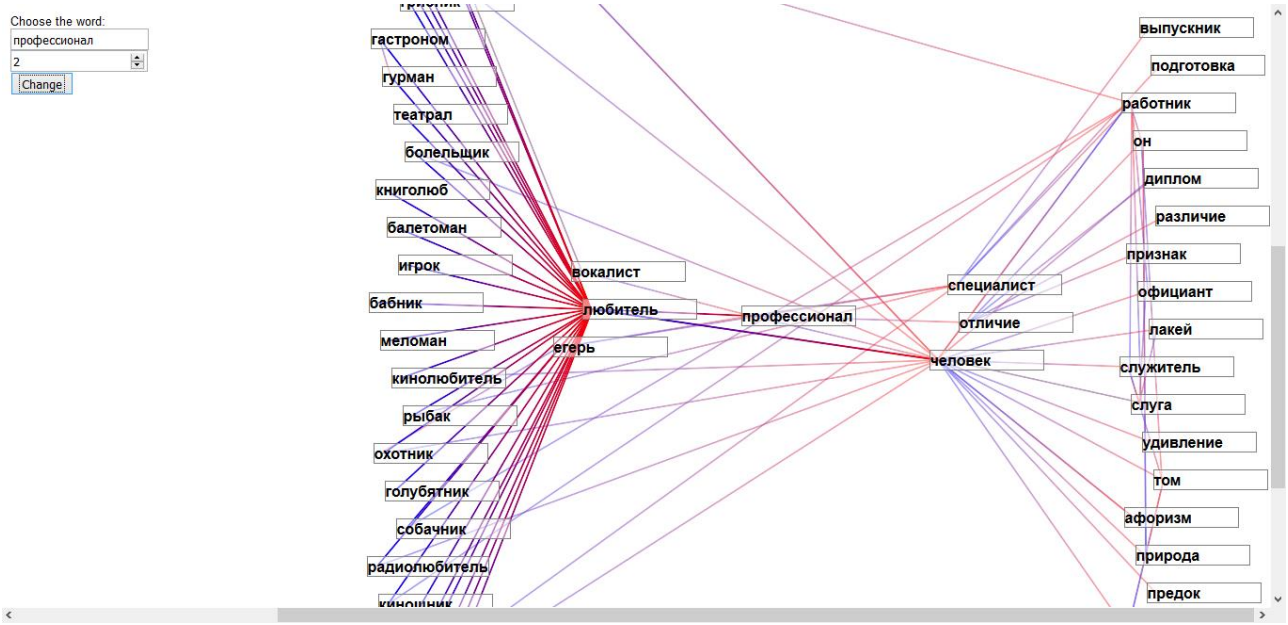



Рис. 3. Скриншот окна программы визуализации дерева связей

Пусть граф содержит $N$ вершин. Каждая вершина, представляющая собой слово, занумерована (например, в алфавитном порядке). Для $i$-й вершины вводится набор натуральных чисел $\{k_i^1,\ldots,k_i^{n_i}\}$, где $k_i^j$ есть номер слова, которое содержится в словарной статье, отвечающей слову с номером $i$. Верхний индекс отвечает порядковой нумерации слов в словарной статье (первое, второе и т.д.). Этот набор означает, что из вершины $i$ выходят $n_i$ ребер, т.е. $n_i$ есть кратность вершины по выходящим связям. Первое ребро соединяет вершину $i$ с вершиной, имеющей номер $k_i^1$, второе – с $k_i^2$ и т.д.

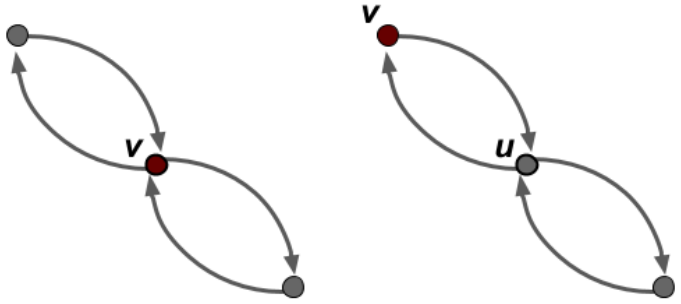
При статистическом анализе онтологического графа словаря использовался следующий алгоритм для подсчета количества циклов

фиксированной длины. В силу сложности вычислительной процедуры мы ограничились подсчетом циклов только из двух, трех и четырех вершин.

Поскольку онтологический граф представлен матрицей смежности $A$, т.е. матрицей, где каждый элемент $A[i][j]$ равен единице, если из $i$ в $j$ есть ребро, и нулю, если между ними нет ребра, то подсчет числа циклов может быть сведен к вычислению следов степеней матрицы смежности. Напомним, что в графе отсутствуют петли, так как в словаре слова не могут определяться через самих себя, поэтому диагональ в матрице смежности нулевая. Теперь заметим, что, например, в случае цикла длины $n$ задача подсчета количества циклов этой длины эквивалентна задаче подсчета путей этой же длины из каждой вершины графа в саму себя. Тогда решение задачи нахождения количества $M_C(k)$ путей, состоящих ровно из $k$ ребер для данной вершины графа, можно выразить формулой:

$$M_C(k) = TrA^k. \qquad (12)$$

Таким образом, просуммировав числа на диагоналях соответствующих матриц, получим количество циклов указанных длин в рассматриваемом графе. Подчеркнем, что интерес представляет задача поиска простых циклов, т.е. таких, которые не проходят через одну вершину дважды. Алгоритм, основанный на формуле (12), примененный для подсчета циклов длины $k > 3$, учитывает также пути, проходящие дважды через одну и ту же вершину:




Для исключения таких циклов учтем два варианта их появления. В первом случае путь проходит дважды через саму интересующую нас вершину, при этом количество таких путей равно квадрату числа таких петель. Во втором случае путь проходит дважды через другую вершину, количество таких путей находится перебором, что повышает вычислительную сложность алгоритма. Приведем результаты расчетов для словаря [2]:

$$M_C(2) = 2886; \quad M_C(3) = 4659; \quad M_C(4) = 49966.$$

**4. Построение онтологического графа слова «предмет»**

В онтологию связей между словами, иерархически объясняющими в рамках конкретного толкового словаря [2] значение некоторого слова, входит примерно 4 тысячи разных слов. Разумеется, детальная визуализация всех связей не поместится на одном листе. Поэтому для пояснения структуры графа мы приведем ниже на рис. 4 фрагменты онтологии, возникающей при анализе словарных статей, порожденных словом «предмет». Подчеркнем, что для

удобства анализа мы сделали некоторую предобработку словаря, исключив из него все союзные слова, а прилагательные, наречия и глаголы заменили образующими их существительными без указания смысловой иерархии внутри одной словарной статьи. Например, слово «местность» определяется в словаре как «1. Какое-нибудь определенное место, пространство, участок земной поверхности» с примерами использования: «гористая местность, открытая местность, занятия на местности», а также как «2. Край (во 2-ом значении), округа» с примерами «дачная местность, в нашей местности». При построении онтологии все поясняющие примеры были исключены. Объясняющие слова были объединены в одну иерархию первого уровня, так что получилось:
«*местность*» = «*определенность, место, пространство, участок, земля, поверхность, край, округа*».

Далее каждое слово иерархии первого уровня представлялось в аналогичном виде, так что по отношению к слову «местность» (в данном примере это есть порождающее слово или слово нулевого иерархического уровня) выстраивалась иерархия второго уровня.

Заметим, что если в словарной статье встречалось само объясняемое слово, оно исключалось из нее. Это сделано потому, что использование в определении (первый уровень) самого определяемого слова (нулевой уровень) является логической ошибкой. Можно, конечно, считать, что смысл подобных слов в данном словаре не определен, но такая позиция неконструктивна. Например, в определении «ткань» = «тканая материя» слово «тканая» исключается, так что остается только «*ткань*» = «*материя*». Таким образом, из фактической онтологии словаря были исключены циклы нулевой длины.

Также отметим, что на данном уровне иерархии каждое слово указывается только один раз безотносительно к тому, сколько раз оно использовалось в рассматриваемой словарной статье. Например, «колода» = «1. Короткое толстое бревно», «2. Предмет, представляющий собою бревно», «3. Комплект игральных карт». Слово «бревно» участвует в объяснении дважды, но учитывается в схеме только один раз. В итоге получается, что «*колода*» = «*краткость, толстота, бревно, предмет, представление, комплект, игра, карта*». Вообще, каждое слово упоминается в онтологии только один раз. Частота его использования отражается числом ориентированных ребер, входящих в соответствующую вершину.

Итак, приведем с учетом сделанных уточнений фрагмент структуры онтологического графа, порожденного словом «предмет». Уровни иерархий будем располагать в параллельных вертикальных столбцах. Связи между словами в соответствии со словарными статьями указываем стрелками. Ниже на рис. 4 показаны только первые три уровня, причем второй и третий – выборочно. Всего бассейн слова «предмет» насчитывает 19 иерархических уровней с общим количеством 3836 разных слов. Совокупная степень вершин графа по входящим ребрам составила 24628, т.е. средняя степень вершины по входящим (и, естественно, по выходящим) ребрам приблизительно равна 6.

Применительно к полному толковому словарю возникает вопрос о связности графа его онтологии: пусто или нет множество пересечения

бассейнов всех слов словаря? Если оно пусто, то существуют непересекающиеся предметные кластеры, специфичные для описания определенной области явлений. Если же есть непустое множество $\Omega = \bigcap_{i=1} B(A_i)$, то в нем лежат все общие классификаторы явлений, связь между которыми устанавливается на уровне множества $V = \bigcup_{k} B(\omega_k)$, где $\omega_k \in \Omega$, по типу переполненного базиса. Вне этого множества располагаются частные классификаторы, в оболочке которых лежат отдельные понятия, объясняемые иерархически через систему общих классификаторов.

По факту оказалось, что кроме более или менее общих классификаторов в бассейн слова «предмет» вошли также и некоторые частные понятия (классификаторы вида или рода), такие как «название животного», «название растения», «вид деятельности», «конкретный предмет». Их доля невелика, но все же они присутствуют в несвойственной им объясняющей категории. Такие слова, видимо, вследствие важности соответствующих объектов в жизни человека, фактически являются аналогами общих классификаторов. Например, в цепочке общих классификаторов появляются видовые названия: одной рыбы, двух насекомых, трех животных и двенадцати растений. Это, соответственно: осетр; оса, пчела; корова, лошадь, куница; береза, виноград, вяз, ива, камелия, конопля, лен, липа, мак, роза, табак, яблоня. Можно провести и более детальный анализ этого бассейна, но на данном этапе это не является целью работы.

В целом бассейн слова «предмет» содержит следующие тематические группы слов, относящиеся: к военным действиям, естественным наукам (физика, химия, математика, геология, анатомия), искусствам (кино, театр, живопись, литература, скульптура), религии, предметам техники и технологии, организации быта, этическим вопросам, а также к явлениям природы.

Следует отметить, что из 19-ти уровней иерархии связей последние 8 имеют относительно случайное содержание и само их число может быть иным – меньше или больше на два-три уровня, в зависимости от выбора синонимов в объясняющих словах предыдущих иерархий.

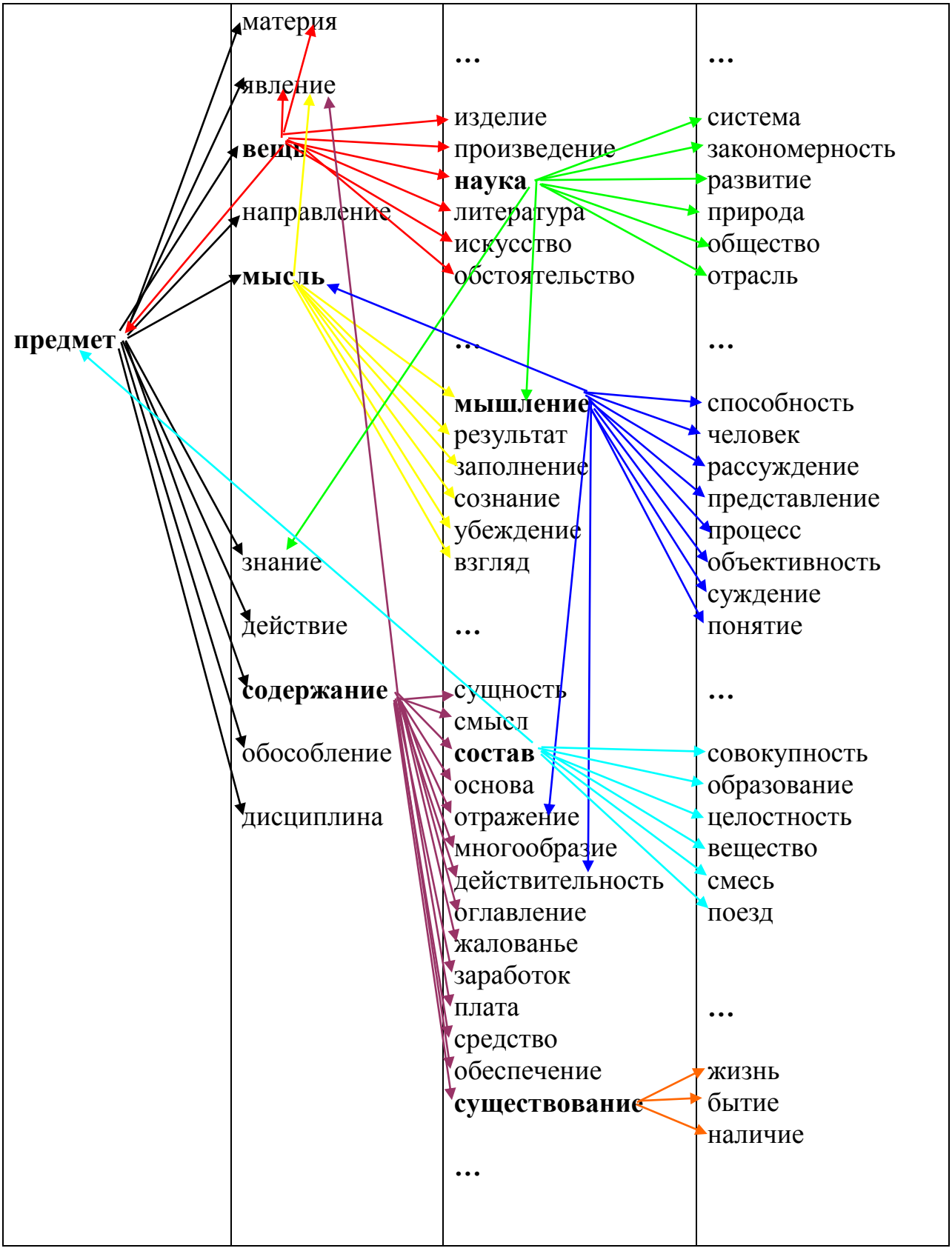



Рис. 4. Фрагмент графа связей, порожденных словом «предмет»
(стрелки разных цветов отвечают выходящим связям разных слов)

На рис. 5 показано распределение слов по уровням. В легенде рис. 5 «число слов» – это полное число выходящих ребер из иерархии данного уровня, «число разных слов» – это число вершин, с которыми соединяются выходящие с данного уровня ребра, а «число вершин» – это число новых слов, появившихся на данном уровне и отсутствовавших на предыдущих уровнях, т.е. слов, не использовавшихся ранее.

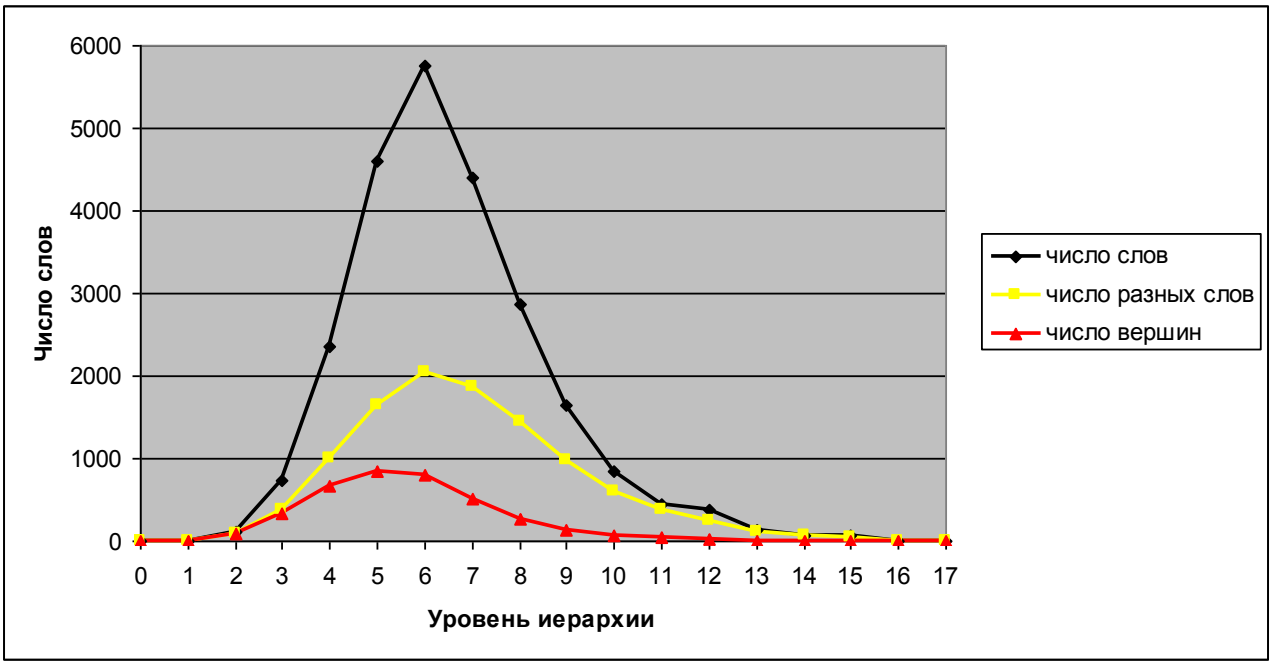



Рис. 5. Распределение числа слов по уровням иерархии бассейна

Наибольшее число слов, а также и наибольшее число ребер появляется на шестом уровне объяснения смысла данного слова, порождающего бассейн. Число вершин, появляющихся на определенном уровне иерархии, имеет максимум несколько раньше – на пятом уровне. Заметим также, что число вершин приблизительно совпадает с числом ребер для первых трех уровней иерархии, после чего число ребер резко возрастает. По-видимому, это означает, что уточнения объяснения смысла исходного слова или фразы на следующих уровнях фактически не происходит, так что толкование возвращается к понятиям, в значительной степени использованным ранее.

На рис. 6 показана доля новых слов (т.е. вершин графа), появляющихся на определенном уровне иерархии, по отношению к полному числу выходящих с этого уровня ребер. С высокой точностью указанная зависимость является степенной с отрицательным показателем степени:

$$K_{new}(t) \approx at^{\gamma}, \quad \gamma \approx -3/2.$$

Это означает, что число связей растет с большей скоростью, чем число вершин.

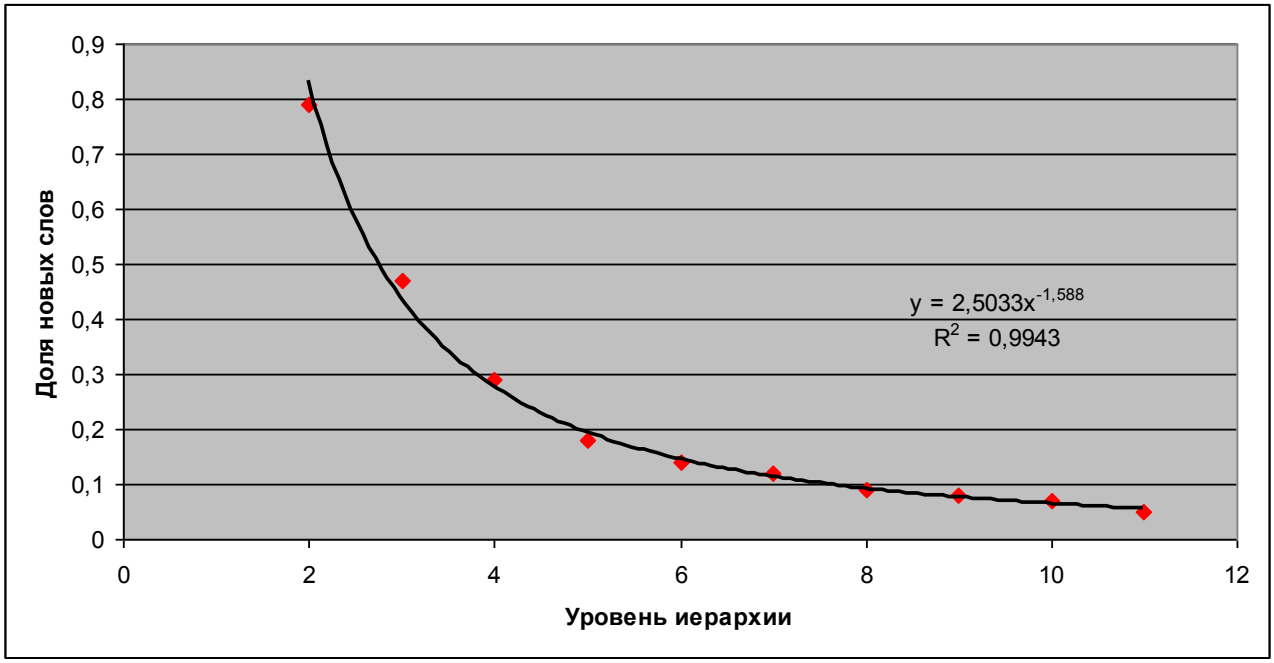



Рис. 6. Доля новых слов в зависимости от уровня иерархии

На рис. 7 показана зависимость функции $V(\mu)$, ранее введенной в (2), от накопленной доли новых вершин. Отметим, что, в отличие от рис. 6, приближение $V(\mu)$ степенной зависимостью имеет заметно меньшую детерминацию ($R^2 = 0{,}88$), чем логарифмическая, для которой $R^2 = 0{,}97$.

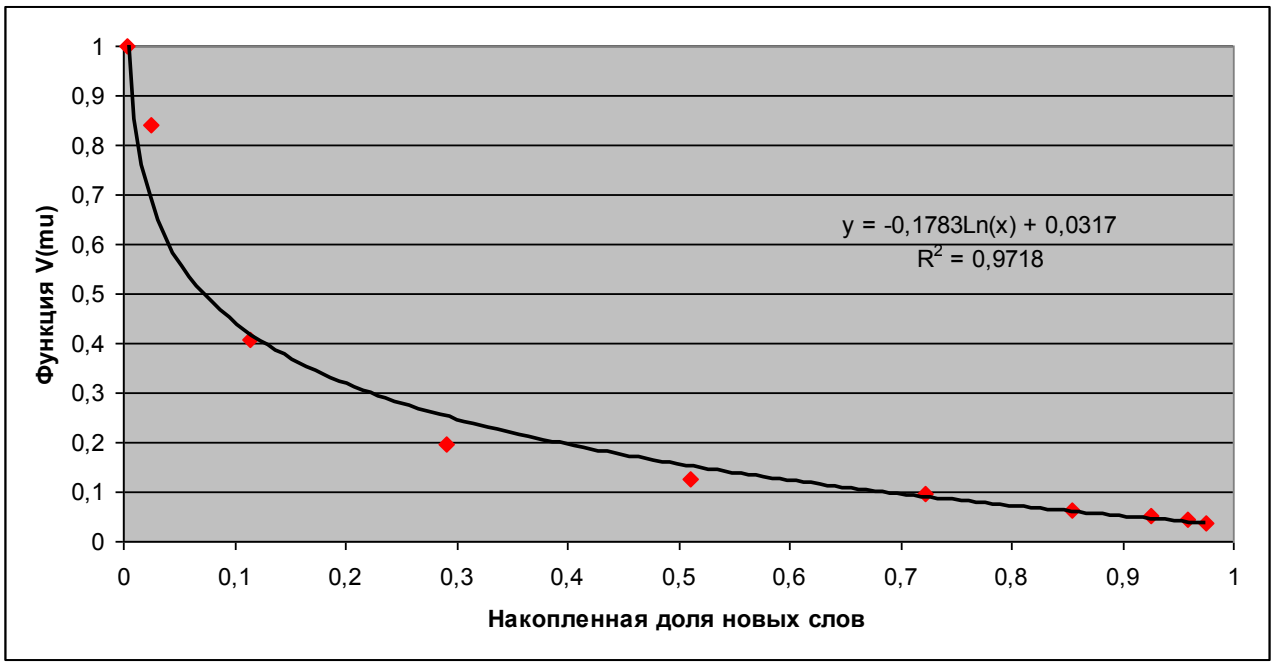



Рис. 7. Зависимость доли новых слов от накопленной доли новых слов

Из построенных 24,5 тыс. связей между словами наибольшую частоту использования имеет слово «часть»: оно встретилось 251 раз. Далее по нисходящей идут слова «человек» (196 раз), «предмет» (166), «отсутствие» (152), «место» (150), «действие» (143), «движение» (118), «состояние» (108),

«жизнь» (100), «форма» (99). Это – первая «десятка» наиболее популярных слов бассейна слова «предмет». Выяснилось также, что доля связей, приходящих к этим словам с каждого уровня иерархии, примерно постоянна для уровней с 3-го по 10-й и не имеет значимой корреляции с номером уровня. Для остальных уровней, не имеющих достаточно большого наполнения словами, указанная доля пренебрежимо мала. Следовательно, модель (2-7), в которой функция $V(\mu)$ считалась приближенно не зависящей от номера уровня иерархии, является допустимой.

Распределение слов бассейна по числу выходящих ребер (т.е. распределение параметра, среднее значение которого равно $\nu$ из (2)), представлено на рис. 8.

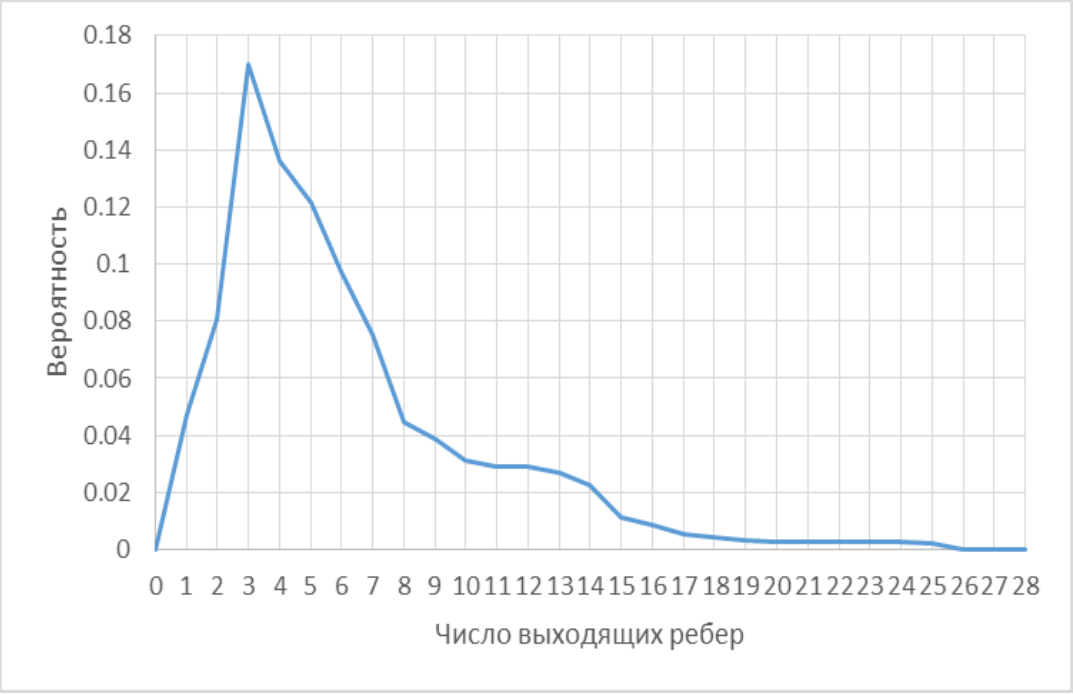



Рис. 8. Распределение числа ребер, выходящих из вершин графа

Как видно, мода ($\nu$ = 3) и среднее значение ($\nu$=6,5) заметно не совпадают. Распределение приближенно описывается функцией $n^{\alpha}\exp(-\beta n)$, аналогичной аппроксимации распределения расстояний между одинаковыми буквами в литературном тексте (см. [17]).

Построенный бассейн слова «предмет» содержит в себе циклы разных длин. Оказалось, что почти для любой пары слов бассейна существует цикл, проходящий через эту пару. Исключение составляют синонимические циклы, образованные парой (тройкой) слов, определяемых одно через другое. Таким образом, все слова, кроме таких тупиковых циклов, являются порождающими бассейн, который тем самым составляет ядро словаря.

**5. Оболочка и граница бассейна**

Добавим теперь к онтологии слова «предмет» другое слово, не содержащееся в его бассейне. Если окажется, что все пути иерархий слова $A'$ продолжаются в бассейне $B(A)$ так, что вне его нет ни одного цикла, то слово $A'$ *будем называть принадлежащим оболочке* $O(A)$ бассейна $B(A)$. Совокупность всех таких слов словаря и образует собственно оболочку $O(A)$. Таковы, например, слова «абажур», «грива», «кучер», «тюльпан» (см. рис. 9), первый-второй уровни иерархии которых состоят из слов $B$(«предмет»).

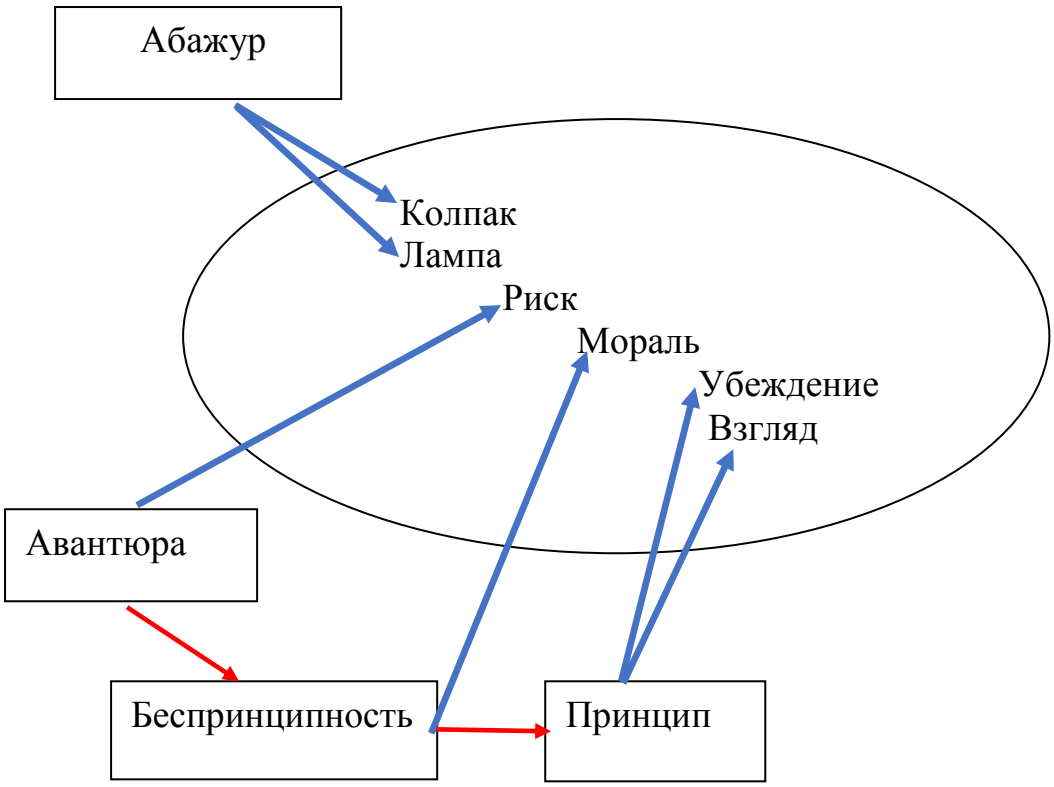



Рис. 9. Фрагмент графа связей, порожденных оболочкой бассейна

Теоретически могла бы возникнуть и такая ситуация, что в иерархиях слова $A'$ содержится несколько циклов до того момента, как в объяснении начнут принимать участие элементы бассейна $B(A)$. Тогда можно выделить часть бассейна $B(A')$, в которой частично реализуется некоторое «объяснение», отличное от онтологии $B(A)$, причем и с бассейном $B(A)$ имеется некоторое непустое пересечение. Возможен также вариант, когда слово $\tilde{A}$ не принадлежит целиком ни $B(A)$, ни $B(A')$, но его иерархии без образования циклов вне указанных бассейнов поглощаются элементами этих бассейнов. Тогда часть слов этих внешних иерархий принадлежит оболочке $O(A)$, а часть – оболочке $O(A')$. *Абсолютной мощностью связи* слова $\tilde{A}$ с бассейном $B(A)$ будем называть число $M(\tilde{A}, A)$ путей, исходящих из $\tilde{A}$ и содержащихся в $O(A)$. Относительной мощностью связи будем называть долю этих путей среди всех

путей, соединяющих слово $\tilde{A}$ с бассейнами других слов. Однако, применительно к рассматриваемому толковому словарю, описанная ситуация не встретилась.

Граничной точкой бассейна будем называть такое слово, принадлежащее бассейну, которое имеет только одно входящее ребро. *Границей* $\Gamma(B(A))$ *бассейна* $B(A)$ будем называть совокупность всех граничных точек этого бассейна. Было выяснено, что граница ядра словаря, т.е. бассейна слова «предмет», состоит из 1360 слов. Это, например, слова: абстракция, агитация, азарт, балет, банк, батарея, ведро, венок, виртуоз, галстук, голод, грусть, ….

Граничные слова, в сущности, представляют частные классификаторы, которые объясняются внутри ядра словаря через более общие понятия.

### 6. Модель базиса ядра словаря

Как выяснилось из анализа бассейна слова «предмет», почти все слова, входящие в него, за исключением тупиковых циклов, являются порождающими этот бассейн. Несмотря на то, что ядро существенно меньше по объему, чем весь словарь, все же оно содержит достаточно много слов (порядка 4 тыс.). Использовать их в качестве базисных «векторов» нецелесообразно, поскольку это будет базис слишком большой размерности, в котором конструктивно определить смысловые нюансы тех или иных предложений окажется затруднительно. Важно также и то, что ранг матрицы смежности не позволяет определить базис ядра. Поскольку слова в словарных статьях можно брать с различным весом, ранг матрицы смежности не имеет большого смысла с лингвистической точки зрения. Базисным элементом словаря является скорее не слово, а цикл, охватывающий определенный набор слов. Количество циклов имеет порядок $N!$, где $N$ – объем ядра, т.е. с практической точки зрения оно бесконечно.

С другой стороны, через элементы ядра определяется в конечном итоге весь словарь. В первом приближении слова можно считать векторами в некотором линейном пространстве, реализацией которого является, например, книга. При этом все элементы ядра словаря связаны между собой циклами. Это означает, что в каждом слове ядра содержится часть (или «проекция») любого другого слова этого же ядра. Такие системы векторов называются переполненными [15, 16]. Для них существует представление так называемых когерентных состояний, разработанное первоначально для описания статистики квантовооптических явлений, но имеющее и более широкое применение. Опишем его в терминах нашей задачи.

Итак, пусть существует в принципе известный, но трудно определяемый конечный (но с учетом вышесказанного можно считать, что счетный) базис смыслов (циклов или их линейных комбинаций), составляющий структуру ядра. Будем обозначать $n$-й элемент этого базиса через $|n\rangle$ и считать, что эти векторы ортонормированы: $\langle k|n\rangle = \delta_{kn}$. Как именно реализован базис, пока не

суть важно. Отдельное же слово будем считать аналогом «когерентного состояния» и обозначать через $|z\rangle$, где $z$ – некоторое число, в общем случае комплексное, которое ставится в соответствие словам ядра, упорядоченным определенным образом, например, в соответствии с частотой встречаемости в ядре. Известно [16], что для таких состояний существует представление вида

$$|z\rangle = \frac{1}{\sqrt{G\left(|z|^2\right)}} \sum_{n=0}^{\infty} \frac{z^n}{\sqrt{Q_n}} |n\rangle, \quad z \in M : \left\{ |z|^2 < R = \lim_{n \to +\infty} \frac{Q_{n+1}}{Q_n} \right\}, \quad G\left(|z|^2\right) = \sum_{n=0}^{\infty} \frac{|z|^2}{Q_n}. \qquad (13)$$

Здесь $Q_n$ – некоторые действительные числа, подлежащие определению, а $R$ – радиус сходимости ряда нормировочной суммы в (13).

Из (13) следует, что

$$\langle n | z \rangle = \frac{z^n}{\sqrt{Q_n G\left(|z|^2\right)}}. \qquad (14)$$

Формула (14) трактуется в том смысле, что величина $\left|\langle n | z \rangle\right|^2$ есть вероятность того, что со словом $|z\rangle$ связан цикл длины *n*.

В круге $M : |z|^2 < R$ может быть выбрана мера $d\mu_z = \rho\left(|z|^2\right) dz d\bar{z}$ такая, что

$$|n\rangle = \frac{1}{\pi\sqrt{Q_n}} \int_M \bar{z}^n |z\rangle d\mu_z, \quad Q_n = \frac{1}{\pi} \int_M \frac{|z|^{2n}}{\sqrt{G\left(|z|^2\right)}} d\mu_z. \qquad (15)$$

Мера $\rho(x)/\sqrt{G(x)}$, моменты которой вычисляются в (15), находится в нашем случае эмпирически: это есть плотность распределения слов ядра по частоте встречаемости, упорядоченных по убыванию частоты (рис. 11), т.е. это распределение слов по числу входящих ребер. Частота убывает приблизительно как $1/\sqrt{n}$, где $n$ - порядковый номер слова.

Радиус $R$ круга $M$ в (15) равен квадрату максимального номера слова в указанном выше распределении, т.е. порядковый номер самого редкого слова в ядре. Таким образом, «мера словаря» определена в кольце $1 \le R \le N^2$.Отсюда вычисляются величины $Q_n$ в (15), после чего известной становится и функция $G(x)$ в (13). В итоге нам становится известной и мера $\rho(x)$.

Теперь мы можем представить неортогональное разбиение единицы, связанное с введенными состояниями, в следующем виде:

$$\hat{I} = \sum_{n=0}^{\infty} |n\rangle\langle n| = \frac{1}{\pi} \int_M |z\rangle\langle z| \sqrt{G\left(|z|^2\right)} \rho\left(|z|^2\right) dz d\bar{z}. \qquad (16)$$

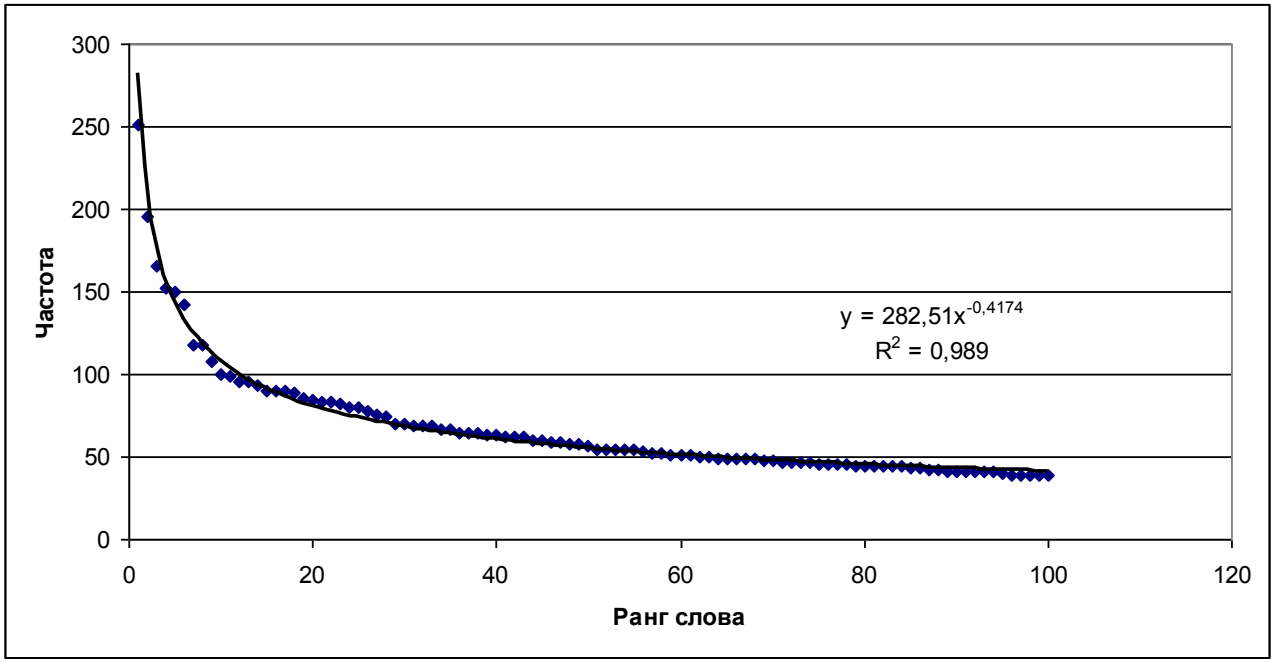



Рис. 11. Фрагмент распределения слов бассейна *B*(«предмет») по частоте встречаемости

Из (13) следует, что если у нас есть два слова, т.е. два когерентных состояния $|z\rangle$ и $|w\rangle$, то проекция одного из них на другое, понимаемая как скалярное произведение $\langle z|w\rangle$, определяется формулой

$$\langle z|w\rangle = \frac{1}{\sqrt{G\left(|z|^2\right)G\left(|w|^2\right)}}\sum_{k,n=0}^{\infty}\frac{\bar{z}^k w^n}{\sqrt{Q_k Q_n}}\langle k|n\rangle = \\ = \frac{1}{\sqrt{G\left(|z|^2\right)G\left(|w|^2\right)}}\sum_{k,n=0}^{\infty}\frac{(\bar{z}w)^n}{Q_n} = \frac{G(\bar{z}w)}{\sqrt{G\left(|z|^2\right)G\left(|w|^2\right)}}. \quad (17)$$

Это выражение можно считать количественной мерой семантической близости. На практике соответствующее значение $G(\bar{z}w)$ естественно вычислять на действительной прямой.

Аналогично определяется и угол между высказываниями, т.к. последние представляют собой выражения вида $|\psi\rangle = \sum_k a_k |z_k\rangle$, $|\varphi\rangle = \sum_l b_l |w_l\rangle$. Близость между этими высказываниями (предложениями, текстами и т.п.) определяется выражением $\langle\varphi|\psi\rangle / \sqrt{\langle\varphi|\varphi\rangle\langle\psi|\psi\rangle}$. Оценка близости, гарантирующая понимание в соответствии с формулой (17), требует специального исследования, которое будет проведено в отдельной работе.

### 7. Результаты анализа и дальнейшие перспективы

В результате проведенного анализа выяснилось, что толковый словарь состоит из ядра, содержащего порядка 4 тыс. слов, а также ближней (25 тыс.) и

дальней (6 тыс.) периферии. Ближняя периферия представляет собой оболочку ядра в терминологии п.5, а дальняя характеризуется тем, что имеет небольшое количество циклов вне ядра. Выявленная структура существенно упрощает анализ связей словарных статей, поскольку для слов ближней периферии устанавливается идеальная онтология ребер, связывающих их с ядром, а дальняя периферия имеет малое число циклов. Поэтому детальный анализ связей достаточно провести только для слов сильно связной компоненты графа (т.е. ядра словаря), что имеет практическую важность.

Также следует сказать, что в анализируемом словаре обнаружились логические ошибки, не позволяющие получить объяснения некоторых используемых терминов. Было найдено четыре типа логических ошибок.

Первая ошибка уже упоминалась выше в п.4 – это наличие циклов нулевой длины, когда определяемое слово составляет основу самой словарной статьи. Например, «пожертвование» определяется как «то, что пожертвовано, дар», слово «дар» определяется как «подарок, пожертвование», а «подарок» есть «вещь, которую дарят».

Вторая ошибка является расширенным вариантом первой и означает наличие так называемых тупиковых циклов, образованных синонимами, которые определяются сами через себя. Так, на каком-то этапе возникает слово «X», которое объясняется как X = A + B, а затем обнаруживается, что A = B + X и B = A + X. В результате ни одно из этих слов не объяснено. Таково упомянутое выше слово «пожертвование». Еще одним примером такого рода является слово «липкость», определяемое как «прилипание, клейкость», а «клейкость», в свою очередь, определяется как «липкость».

Третья ошибка – это использование в качестве объяснения отрицание отрицания, т.е. некоторое слово X определяется с использованием конструкции «не-не-X». Так, «север» определяется как «одна из четырех стран света, направление, противоположное югу». Излишне даже и упоминать, что «юг» определяется как «направление, противоположное северу».

Четвертая ошибка представляет собой запутывание второй ошибки и состоит в том, что два разных слова раскрываются через одинаковую словарную статью, но не имеют прямой связи друг с другом. В результате образуется множество ложных циклов, означающих только то, что X = Y. Таковы, например, статьи «царь», «император», «король», которые определяются как «титул монарха в некоторых странах», но между собой не связаны синонимическим отношением. Другой пример такого рода: слово «содержание» имеет кроме «оглавления» еще и смысл «жалованье, заработок, плата»; при этом «жалованье», как и «плата», определяются одинаково – это «денежное вознаграждение за работу».

Разумеется, будучи полностью циклической онтологией, словарь ничего не может определить в строгом смысле этого слова, так как для этого в основе должны лежать некие неопределяемые слова, через которые будут выражаться все остальные. Но цель словаря не в этом, он предназначен для установления причинно-следственных связей между понятиями по принципу присутствия каждого слова в раскрытии смысла каждого слова. Наличие тупиковых циклов

нарушает конструкцию так называемого «переполненного базиса», которым в удобно моделировать идеальное ядро словаря. Один из вопросов, требующих специального исследования, состоит в том, насколько можно скорректировать определения словаря, чтобы в нем не было прямых логических ошибок, и насколько ядро чувствительно к выбору объясняющих слов.

Также выяснилось, что в словаре содержатся не вполне корректные определения некоторых понятий, имеющих наряду с общим и узко специальное использование. В результате формируется искаженный граф связей между словами. То, какие именно слова и в каком количестве войдут в ядро словаря, зависит в том числе и от точности даваемых определений. Например, статья «квадратура» дается в контексте выражения «квадратура круга» и обозначает следующее: «неразрешимая задача превращения круга в равновеликий квадрат». Это, строго говоря, неверно, поскольку данная задача не вообще «неразрешима», а не имеет решения с применением только циркуля и линейки. Используемое слово «превращение» еще более запутывает и не слишком подходит для объяснения математического термина, причем в рамках самого словаря оно означает «придание иного вида», что тем более не позволяет понять проблему неразрешимости: круг ведь можно растянуть и деформировать до состояния квадрата. Другой пример – статья для слова «предел». Она имеет три подраздела: два общих (страна, местность; крайняя степень чего-либо в контексте «дойти до предела возможностей») и один математический. Последний звучит так: «пространственная или временная граница чего-либо». То, что это самое «что-либо» должно присутствовать в любой двусторонней окрестности упомянутой «границы», из данного определения не следует, но тогда это и не предел в его строгом понимании.

К группе не вполне корректных определений относятся также определения цветовой окраски. В ядро словаря входят семь цветов, но не все они составляют «радугу». Используемые цвета следующие: красный, желтый, синий, зеленый, а также белый, черный, серый. Они определены не с физической точки зрения (что для языковедения, наверное, правильно), но и не так, как остальные слова ядра словаря, т.е. не с помощью общих классификаторов. Они определены через конкретные предметы, являющиеся частными классификаторами. Так, красный – это цвет крови, мака и ягод земляники; желтый – яичного желтка; синий – безоблачного неба; зеленый – травы; белый – серебра; черный – сажи; серый – золы. В результате появляются нетипичные связи между словами, что, возможно, приводит к дополнительным уровням иерархии.

Конечно, эти замечания можно считать а) несущественными и б) легко поправимыми. Но в этой связи возникает задача устойчивости глубины онтологии в зависимости от выбора формулировок, что, впрочем, выходит за рамки данного исследования.

**Литература**

1. Бова В.В., Кравченко Д.Ю., Лещанов Д.В., Новиков А.А. Компьютерная онтология: задачи и методология построения // Информатика, вычислительная техника и инженерное образование, 2014. – № 4.
2. Ожегов С.И., Шведова Н.Ю. Толковый словарь русского языка. – М.: Русский язык, 1989.
3. Guriano N. Understanding, Building, and Using Ontologies / A Commentary to «Using Explicit Ontologies in KBS Development» // International Journal of Human and Computer Studies, 1997. V. 46. № 2/3.
4. Котеленко С.А. Формальное описание онтологий на основе нечеткой гиперграфовой модели данных // Изв. ЮФУ, Технические науки, 2005. – № 6. С. 138-145.
5. Велихов П.Е. Меры семантической близости статей Википедии и их применение к обработке текстов // Информационные технологии и вычислительные системы, 2009. – № 1. С. 23-37.
6. Leicht, E. A. and Holme, Petter and Newman, M. E. J., Vertex similarity in networks // Phys. Rev. E, 73:026120, 2006.
7. Гаврилов С.В., Кирильченко А.А., Кугушев Е.И., Платонов А.К. Программный комплекс для создания и анализа семантических сетей, описывающих системы понятий // Препринты ИПМ им. М.В.Келдыша, 1986. – № 140.
8. Sho Yoshida, Hiroaki Tsurumaru, Tooru Hitaka. Man-assisted machine construction of a semantic dictionary for natural language processing / COLING '82 Proceedings of the 9th conference on Computational linguistics – V. 1, pp. 419-424
9. Segalovich, I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine / Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications. MLMTA’03. - Las Vegas.- 2003. - P. 273-280.
10. Porter M. An algorithm for suffix stripping. Program 14.3 (1980). P. 130-137.
11. Larsen B., Chinatsu A. Fast and effective text inning using linear-time document clustering. In Proc. KDD, 1999. P. 16–22.
12. Kukich K. Techniques for automatically correcting words in text. ACM Computing Surveys, 1992. V. 24(4). P. 377–439.
13. Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein. Introduction to Algorithms, 3rd edition. The MIT Press, 2009.
14. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.
15. Березин Ф.А. Метод вторичного квантования. – М.: Наука, 1986.
16. Орлов Ю.Н. Основы квантования вырожденных динамических систем. – М.: МФТИ, 2004.
17. Орлов Ю.Н., Осминин К.П. Методы статистического анализа литературных текстов. – М.: Эдиториал УРСС/Книжный дом «ЛИБРОКОМ», 2012.