# Алгоритм выделения первичной семантической структуры в коллекции определений

Тузикова А.В.
Факультет вычислительной математики и кибернетики,
Московский государственный университет имени М.В. Ломоносова,
г. Москва, Россия
Email: kabaylova@yandex.ru

Аннотация—В работе рассматривается возможность частичной автоматизации процесса формирования терминологических сетей. Предлагается подход к выявлению семантической структуры, основанный на использовании методов гибридной кластеризации и выделения семантических связей. Работоспособность метода подтверждается на модельной коллекции определений.

Ключевые слова—терминологическая сеть, семантическая структура, семантическая группа, семантическая связь.

## I. Введение

Динамичное развитие современного общества приводит к росту объемов информации. Потребность структурировать знания, которые накоплены человечеством за долгий период его существования, ощущается весьма остро. Только упорядоченное представление знаний позволит сформировать целостную картину мира, объективно отражающую действительность. Модель структуризации на основе семантических сетей является достаточно гибкой, она позволяет подстраиваться под текущие нужды, не теряя при этом качества представления знаний. Проект “Универсальное терминологическое пространство” (УТП) [5] опирается на особый класс семантических сетей, именуемых терминологическими сетями.

Развитие УТП обеспечивается ручным трудом экспертов-редакторов. Качественное расширение терминологической сети является отправной точкой для задачи автоматизации этой деятельности.

## II. Введение в терминологические сети

Под терминологической сетью [6] будем понимать семантическую сеть [1], состоящую из определений терминов и бинарных отношений между ними (из вершин и дуг соответственно). Вершина сети фактически является структурой данных, содержащей исчерпывающую информацию о термине, а также о связанных с ним вершинах. В качестве ребер используется лишь два допустимых типа отношений: “это-есть” и “относится-к”. Другие типы связей полагаются сводящимися к используемым или учтенными в структуре вершин сети (например, отношение синонимии).

Добавление новых коллекций определений в существующую структуру предполагает внимательное изучение и обработку редактором сети новой информации. Каждое определение в коллекции представлено в формате: <термин> – <определение>. При взаимодействии с коллекциями значительных размеров приходится говорить о возрастании сложности обработки неструктурированных данных. Для облегчения деятельности редактора предлагается алгоритм выявления первичной семантической структуры коллекции определений, основанный на разделении этапов выявления семантических групп и семантических связей.

## III. Понятие семантической группы

В лингвистике для изучения системной организации языка используется понятие “семантическое поле”, получившее широкое распространение в начале 20-го века [8]. Теория поля представляет собой интерес, поскольку позволяет отойти от рассмотрения уровневой модели языка.

Семантическое поле определяется по-разному, например, как «совокупность языковых (главным образом лексических) единиц, объединённых общностью содержания (иногда также общностью формальных показателей) и отражающих понятийное, предметное или функциональное сходство обозначаемых явлений» [4]. Фактически, семантическое поле является системой взаимосвязей предопределенных типов, представляющей из себя семантическую сеть. Обязательным требованием к семантическому полю является некоторая общность входящих в него единиц, обязательная полнота или взаимодействие с остальной языковой системой не требуются.

Во избежание недоразумений, связанных с отсутствием точного определения, отойдем от использования термина поле. Основополагающими факторами для принятия такого решения являются: неприменимость методов построения семантических полей в силу своих особенностей в рамках поставленной задачи, а также относительный характер общности по содержанию единиц, входящих в коллекцию.

Далее под термином “семантическая группа” будем понимать совокупность семантически близких языковых единиц, относящихся к одной предметной области. Такой подход к определению позволяет рассматривать обрабатываемую коллекцию как объединение непересекающихся, но связанных семантических групп.

## IV. Автоматическое выявление первичной семантической структуры

При выявлении первичной семантической структуры предлагается использовать двухэтапный метод последовательного выявления семантических групп и возможных связей с целью последующего включения в существующую терминологическую сеть. Жесткая последовательность этапов обусловлена использованием результатов применения алгоритма выявления групп при поиске связей.

Технология выявления семантической структуры строится следующим образом:

Этап 1. Выявление семантических групп.
1) Первичная кластеризация.
2) Интуитивная кластеризация.

Этап 2. Выявление семантических связей.
1) Определение внутренних связей.
2) Определение межгрупповых связей.
3) Добавление кандидатов.

Рассмотрим каждый из этапов в отдельности.

### A. Описание этапа выявления семантических групп

В основу алгоритма выявления семантических групп положено предположение о возможности выделения совокупностей объектов, образующих некоторую смысловую целостность путем использования методов кластеризации.

Коллекция определений, предлагаемая для обработки, представляет собой набор предложений, что позволяет использовать модель “мешок слов”. Учитывается частота употребления слов, но не их порядок или грамматические формы.

Алгоритмы кластеризации позволяют получить группы определений, в которых употребляется большое количество общих терминов. Таким образом, объекты внутри одного кластера действительно оказываются близкими друг к другу. Однако в целях локализации еще более близких по смыслу понятий предлагается использовать метод «интуитивной кластеризации», основанный на предположении о схожих названиях родственных понятий. Например: акция, акционер, акционерное общество.

Метод интуитивной кластеризации применяется только к определяемым терминам, что позволяет ожидать получения корректных результатов. Для каждого слова в составе термина рассматриваются лишь несколько первых букв, количество которых является постоянной величиной, установленной экспериментально. На основании полученных n-грамм выявляются близкие понятия, объединяемые в искомые совокупности.

Таким образом, задача структуризации входной коллекции определений сводится к двум задачам: задаче выявления семантических групп и задаче поиска семантических связей между формируемыми понятийными вершинами.

Предложенный метод выделения семантически близких групп определений позволяет облегчить процесс ручной коррекции непосредственно перед внедрением новых данных в существующую терминологическую сеть.

Резюмируя приведенные рассуждения приведем следующий алгоритм.

### A.A. Алгоритм выявления семантических групп

Входными данными для работы алгоритма является заданная пользователем коллекция определений.

Выходными данными алгоритма являются группы семантически близких определений, обрабатываемые в рамках второго этапа выявления семантической структуры.

Алгоритм выявления групп состоит из двух этапов.

Этап 1. [Первичная кластеризация]. Определить границы смысловых совокупностей обрабатываемых определений с использованием классических методов кластеризации. При использовании иерархических методов кластеризации – остановить процесс выявления групп при достижении ожидаемого количества на текущем этапе.

Этап 2. [Интуитивная кластеризация]. Для каждого объединения, полученного на первом этапе, применить алгоритм интуитивной кластеризации к левым частям определений.

Алгоритм завершает работу после того, как все совокупности объектов, выявленные на этапе первичной кластеризации будут разделены на непересекающиеся семантические группы. Результатом применения алгоритма к входной коллекции определений является набор групп, число которых индивидуально для каждой конкретной коллекции.

По завершении выделения определений, составляющих семантические группы, результаты предоставляются алгоритму выявления семантических связей.

### B. Описание этапа выявления семантических связей

Для выявления семантических связей предлагается использовать метод, основанный на определении псевдорасстояний между предполагаемыми вершинами терминологической сети [7].

Алгоритм основан на предположении о наличии в тексте определения достаточной информации для определения его семантической близости с другими определениями. Поскольку все данные представлены в виде строк, расстояния между определениями можно определять с использованием метрик редактирования [2]. Однако, рассматривая отдельные слова, а не определения в целом, можно ограничиться только операциями замены и вставки-удаления символов, то есть использовать метрику Левенштейна [3].

### B.A. Общая идея алгоритма выявления связей

Входными данными для работы алгоритма являются: обрабатываемое определение и набор определений, с которыми устанавливаются связи.

Выходными данными алгоритма являются элементы входного набора, с которыми могут быть установлены бинарные отношения одного из предопределенных структурой терминологической сети типов.

Алгоритм выявления связей является двухэтапным.

Этап 1. [Определение расстояний]. Определить псевдорасстояния между обрабатываемым определением и каждым определением из входного набора с использованием метрики Левенштейна.

Этап 2. [Добавление кандидатов]. Добавить подходящие элементы в итоговое множество кандидатов на добавление связей. Если множество заполнено, производить замещение худшего из текущих вариантов.

Результатом работы является совокупность определений, потенциально связанных с обрабатываемым.

### B.B. Алгоритм выявления семантических связей

Входными данными для работы алгоритма являются выделенные семантические группы.

Алгоритм выявления связей применяется для формирования структуры входной коллекции определений в три этапа.

Этап 1. [Определение внутренних связей]. Определить потенциальный набор связанных определений внутри отдельных семантических групп.

Этап 2. [Определение межгрупповых связей]. Определить потенциальный набор связанных определений, относящихся к различным семантическим группам. Для каждого определения рассмотреть в качестве входного набора всю исходную коллекцию определений за исключением собственной группы.

Этап 3. [Добавление кандидатов]. Уменьшить значимость межгрупповых связей на величину, соответствующую одному несущественному сравнению. Сформировать итоговое множество.

Результатом применения алгоритма является совокупность пар элементов, которые могут быть связаны. Полученная совокупность предлагается редактору для выбора корректных связей. Выделение семантических групп обеспечивает обозримость обрабатываемых данных, чем существенно снижает нагрузку на редактора.

По результатам взаимодействия редактора с системой терминологическая сеть оказывается расширенной за счет сформированных понятийных вершин и связей между этими вершинами.

## V. Применение метода выявления семантической структуры к модельной предметной области

В главе проверяется работоспособность метода выявления семантической структуры на модельной коллекции определений и приводится анализ полученных результатов.

### A. Выявление семантических групп

Для проверки работы метода выявления семантических групп проведен ряд экспериментов. Для получения результатов использовалась коллекция из 800 определений. К этим данным применены как алгоритмы кластеризации в исходном виде, так и гибридный алгоритм для различных первых этапов.

Оценка результатов экспериментов проводилась по формуле:

$$Res = \frac{\sum_{i=0}^{k} c_i/n_i}{k} \times 100\% \quad (1)$$

Здесь $k$ – число кластеров, $c_i$ и $n_i$ – число правильно определенных элементов соответствующего кластера и общее число элементов, отнесенных к нему, соответственно.

### A.A. Классические алгоритмы кластеризации

Для модельной коллекции проведено выявление семантических групп с использованием классических методов кластеризации. В таблице I приведены полученные оценки результатов для метода k-средних.

В предложенной таблице столбец “Определено” содержит количество объектов, верно отнесенных к соответствующим кластерам, “Эксперимент” – номер эксперимента, для которого приведены данные, “Результат” – оценку качества приведенного метода, выраженную в процентах.

Таблица I. Оценка результатов метода k-средних для выделения семантических групп

| Эксперимент | Определено | Результат |
|---|---|---|
| 1 | 491 | 70.77% |
| 2 | 457 | 55.89% |
| 10 | 480 | 62.71% |
| Среднее за 20 | 487 | 64.25% |

Для метода k-средних наблюдается большой процент ошибки, что говорит о неверно выбранном числе кластеров. Таким образом, следует выделять семантические группы в исходной совокупности с учетом не только содержательной близости.

Метод средней связи показал себя лучше, чем метод k-средних: верно распределены по группам 582 объекта, оценка качества составляет 76.30%. Основным преимуществом метода с теоретической точки зрения является учет всех объектов кластера, что снижает влияние элементов, плохо вписывающихся в кластер.

### A.B. Гибридный алгоритм кластеризации

Оценки результатов выделения семантических групп с использованием гибридной модели на основе метода k-средних приведены в таблице II. В таблице III приведены результаты применения указанного алгоритма на первом этапе для выделения крупных кластеров.

Таблица II. Оценка результатов гибридной кластеризации для выделения семантических групп на основе метода k-средних

| Эксперимент | Определено | Результат |
|---|---|---|
| 1 | 629 | 78.02% |
| 2 | 482 | 60.76% |
| 10 | 549 | 67.59% |
| Среднее за 20 | 576 | 74.69% |

Таблица III. Оценка результатов метода k-средних на первом этапе гибридной кластеризации

| Эксперимент | Определено | Результат |
|---|---|---|
| 1 | 686 | 82.95% |
| 2 | 503 | 64.80% |
| 10 | 583 | 70.71% |
| Среднее за 20 | 594 | 71.92% |

Данные, представленные в Таблицах I, II и III, позволяют судить о том, что гибридный алгоритм, основанный на методе k-средних, позволяет получить более качественные результаты, чем использование метода k-средних для получения желаемого количества семантических групп из исходного набора данных.

Для метода кластеризации на основе средней связи ситуация аналогична. В случае гибридной кластеризации верно обработаны 606 объектов, полученный результат – 77.34%. Для метода средней связи на первом этапе гибридной кластеризации эти показатели составляют 651 объект и 84.42%. Точность выявления семантических групп оказывается выше, чем при использовании классического метода.

Более высокие показатели при выделении меньшего числа кластеров на первом этапе работы алгоритма обусловлены разделением входной коллекции на смысловые группы, с которыми ведется дальнейшее взаимодействие в рамках второго этапа гибридной модели, положенной в основу метода выявления семантических групп. Следует отметить, что, несмотря на падение точности, представляется невозможным завершить процесс выявления семантических групп после первого этапа гибридной модели, поскольку в этот момент группы определений все еще имеют значительные размеры.

Таким образом, выявление семантических групп во входной коллекции определений с целью последующего выявления связей для включения в терминологическую сеть предпочтительно проводить с использованием описанного алгоритма (глава IV, раздел A).

### B. Описание результатов

При последовательном применении алгоритмов выявления семантических групп и семантических связей наблюдается некоторое падение точности; установлено, что формирование корректных вершин терминологической сети и соответствующих им связей предопределенных типов только на основании выявленных данных возможно в 71% случаев. Это явление вызвано избыточностью предлагаемых редактору вариантов потенциально связанных вершин по завершении работы алгоритмов – основным фактором является выявление не только внутренних связей в каждой семантической группе, но и внешних связей относительно своей группы для каждого определения.

Таким образом, результат работы алгоритма выявления семантической структуры с точки зрения выявленных связей практически идентичен теоретическим результатам применения метода к неразделенной на группы входной коллекции определений. Однако выявление семантических групп существенно повышает локальность данных и их наглядность для редактора терминологической сети.

## VI. Заключение

Предложен метод предварительной структуризации данных с целью их последующего включения в терминологическую сеть, основанный на последовательном использовании методов кластеризации для выявления семантических групп и метода выявления семантических связей.

Использование метода гибридной кластеризации (глава IV, раздел A) позволяет локализовать определения, потенциально связанные между собой.

Метод выявления связей (глава IV, раздел B) дает возможность свести процесс построения семантической сети на основании коллекции определений к выбору типа и объекта связи.

Описанный метод выявления первичной семантической структуры (глава IV) демонстрирует приемлемую точность и позволяет за счет частичной автоматизации повысить эффективность работы редактора терминологической сети и значительно облегчить его деятельность при расширении проекта за счет относительно небольших объемов данных и предоставления вариантов возможных связей.

Предложенный подход позволяет решать задачи автоматизации построения терминологических сетей с использованием дедуктивного машинного обучения.

## Список литературы


[1] John F. Sowa Semantic Networks / Encyclopedia of Artificial Intelligence, second edition, Wiley, New York, 1992.

[2] Деза Е.И., Деза М.-М. Энциклопедический словарь расстояний / Елена Деза, Мишель-Мари Деза – пер. с англ. В.И. Сычева; Моск. гос. пед. ун-т ; Нормальная высш. шк., Париж. – М.: Наука, 2008. – с. 178-186.

[3] Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады Академии Наук СССР, т. 163, №4 – М.: Наука, 1965. – с. 845–848.

[4] Лингвистический энциклопедический словарь / гл. ред. В.Н. Ярцева. – М.: Новый Диск / ДиректМедиа, 2000. – 746 с. – (ЛЭС). с.180-181

[5] Мальковский М.Г., Соловьев С.Ю. Универсальное терминологическое пространство. // Труды Международного семинара Диалог'2002 "Компьютерная лингвистика и интеллектуальные технологии т.1. М.: Наука, 2002, с. 266-270.

[6] Мальковский М.Г. Терминологические сети / М.Г.Мальковский, С.Ю.Соловьев // OSTIS-2012. Материалы конференции. С. 77-82.

[7] Тузикова А.В. Алгоритм выделения бинарных отношений в терминологической сети // OSTIS-2016. Материалы конференции. Минск: БГУИР, 2016, с. 187.

[8] Щур Г.С. Теория поля в лингвистике. / М.: Наука, 1974. - 256с., с.34.