О кластеризации терминологических сетей

Дмитриев К.А.
Факультет вычислительной математики и кибернетики,
Московский государственный университет
имени М.В. Ломоносова,
г. Москва, Россия
Email: dmitriev_ka@akticom.ru

Аннотация—В работе рассматривается задача кластеризации терминологической сети, а также способ ее решения; предлагается модифицированный алгоритм выделения сообществ в графе, позволяющий учитывать особенности терминосистем. Кроме того, в работе описывается способ оценки качества кластеризации сети путем вычисления модулярности и приводится описание метода, позволяющего находить центры кластеров терминологической сети.

Ключевые слова—терминологическая сеть, сообщество, кластер, понятие, модулярность.

I. Введение

Каждый термин, помимо описания некоторого понятия также задает его отношения с другими понятиями в пределах определенной предметной области, образуя таким образом её терминосистему [1]. Структура терминосистем задается совокупностью семантических связей, допускающих объединение в единую терминологическую сеть, представляющую собой ориентированный граф, узлы которого соответствуют терминам, а дуги – бинарным отношениям из допустимого набора [1].

С увеличением числа интегрированных в терминологическую сеть терминосистем возникает проблема навигации, вызванная, в конечном счете, отсутствием разбиения сети на кластеры. Стоит отметить, что кластеризация имеет смысл не только для всей терминологической сети, но и для отдельных ее частей. Фактически кластеризация всей терминологической сети сводится к восстановлению составляющих ее терминосистем.

С точки зрения машинного обучения задача кластеризации терминологической сети аналогична задаче выделения сообществ в социальном графе. Общепринятого формального определения «сообщество» нет [2]. Для данной работы достаточно интуитивного понятия: «сообщество» – это группа вершин сети, узлы которой связаны друг с другом значительно теснее, чем с остальными вершинами.

Выделение сообществ является актуальной задачей в прикладной теории графов. В настоящее время известно множество алгоритмов кластеризации, или поиска сообществ в графе, используемых в различных областях науки, таких, как физика, биология, информатика, прикладная математика и социология. Однако, ни один из этих алгоритмов не учитывает особенности терминологических сетей и не может быть применен для них в явном виде [2], [3], [4]. Тем не менее, указанная проблема может быть решена посредством модификации алгоритмов поиска сообществ определенным образом.

В работе рассматривается модификация алгоритма Fast unfolding of communities in large networks [5] для применения к задаче кластеризации терминологической сети.

II. Терминологические сети

Зафиксируем некоторые особенности терминологической сети, а также принимаемые предпосылки.

Рассмотрим терминологическую сеть $G = (V, E)$, где V - множество вершин-терминов, а E - множество ребер (экземпляров бинарных отношений) двух видов «это-есть» и «относится-к».

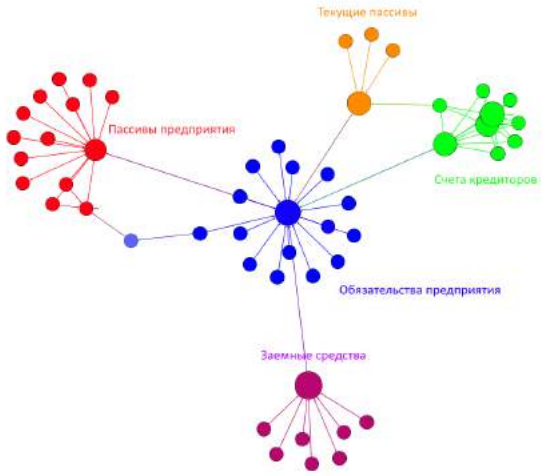

Рис. 1. Пример фрагмента терминологической сети

Ребро (u, v) вида «это-есть» - родовидовое отношение, в котором: v - родовое понятие для u, а u - подвид понятия v.

Ребро (u, v) вида «относится-к» - бинарное отношение, в котором: v является областью применения для u, а u выступает аксессуаром для v. Содержательно u может быть частью, свойством или аспектом v, u может

так же выступать в качестве фигуранта в определении понятия v и т.д. [6].

Вершина v, в которую входит хотя бы одно ребро, т.е $deg^+(v) > 0$ называется понятийной.

Стандартная задача поиска сообществ в графе имеет две постановки: с перекрывающимися сообществами и без перекрывающихся сообществ. В первом случае разным кластерам G' и G'' разрешается иметь одинаковые вершины, во втором случае любая вершина может принадлежать только одному кластеру.

Вообще говоря, исходя из определения терминологической сети, можно сделать вывод о том, что одному узлу не запрещается иметь более одной дуги с узлами, которые могут находиться в разных терминосистемах, поскольку в силу своей природы некоторые понятия могут принадлежать одновременно нескольким предметным областям. Так, например, термин «андеррайтер» относится одновременно к страхованию и биржевому делу. Ввиду того, что принадлежность нескольким предметным областям встречается редко, в работе используется допущение о том, что сообщества не перекрываются и такая ситуация невозможна, то есть, термин «андеррайтер» в данном случае будет отнесен только к одному из кластеров.

Будем также полагать, что решаемая задача состоит в отыскании кластеров внутри сети, каждый из которых представляет собой одну и только одну компоненту слабой связности, т.е. такой граф, что, при игнорировании направления дуг которого он является связным, в противном случае может быть разделен.

В работе понятия «кластер» и «сообщество» считаются эквивалентными, и в введенных терминах представляют собой подграф $G' = (V', E')$ исходного графа $G = (V, E)$, где V' — некоторое подмножество V и E' — подмножество всех ребер графа G, концевые вершины которых входят в V'.

III. Кластеризация

Для кластеризации терминологической сети предлагается использовать алгоритм состоящий из трех шагов:

1) выделить кластеры с помощью эвристического алгоритма, оптимизирующего модулярность [5];
2) вычислить для каждого кластера i величины

$$K_i = \frac{P_i}{N_i},$$

где P_i - число понятийных узлов в i-м кластере, N_i - общее число вершин в i-м кластере;
3) выделить центры кластеров, для которых величина $K_i > 0.3$ и разделить их на более мелкие;

Рассмотрим шаги агоритма более подробно.

A. Первый шаг

Основная идея алгоритма кластеризации состоит в максимизации значения функционала Q, называемого модулярностью.

Q есть скалярная величина из отрезка $[-1, 1]$, вычисляющаяся по формуле:

$$Q = \frac{1}{2m} \sum_{i,j} (A_{ij} - \frac{d_i d_j}{2m}) \sigma(c_i, c_j),$$

где

$$\sigma(c_i, c_j) = \begin{cases} 1 & c_i = c_j \\ 0 & c_i \neq c_j \end{cases},$$

A — матрица смежности графа, A_{ij} — i, j элемент матрицы, d_i — степень i вершины графа, c_i — номер кластера, к которому относится вершина, m — общее количество ребер в графе.

Функционал Q был предложен в [4] и в настоящее время вычисление значения функционала является одним из популярных способов оценки качества решения задачи кластеризации графа [2].

Использование функционала Q сводит задачу разбиения терминологической сети на кластеры к поиску для каждой вершины номера c_i, при которых значение функционала достигает максимума.

С содержательной точки зрения, модулярность есть разность между долями ребер внутри сообщества и ожидаемой доли связей, получаемой если бы все ребра терминологической сети были бы размещены случайным образом при условии сохранения степеней вершин исходного графа.

Для алгоритмов выделения сообществ, основанных на модулярности, известна проблема выделения малых кластеров в больших графах, именуемая в дальнейшем проблемой масштабируемости: из-за того, что в случайном графе ребра могут соединять любые вершины, ожидаемое число ребер между двумя сообществами может принимать значения меньше единицы. В такой ситуации наличие ребра между двумя кластерами будет интерпретироваться как признак сильной корреляции между двумя кластерами и в процессе оптимизации модулярности приводить к их слиянию.

Таким образом, даже слабосвязанные между собой полные графы, которые имеют максимально возможную плотность внутренних ребер и представляют собой легко выделяемые сообщества, будут объединены в один кластер в ходе максимизации модулярности [7].

Один из возможных путей решения проблемы масштабируемости - модификация формулы для вычисления модулярности путем добавления параметра масштаба [8]. Другой способ - использование в алгоритме особенностей кластеризуемого графа.

B. Второй шаг

Для решения проблемы масштабируемости необходимо выделить те кластеры, в которых после первого

шага выполнения алгоритма она могла потенциально возникнуть.

С этой целью для каждого полученного на первом шаге кластера вычисляется значение K_i, представляющее собой отношение числа понятийных вершин к общему числу понятий в кластере. Отметим, что $K_i \in [0, 1]$, причем $K_i = 0$ только в тривиальном случае, когда кластер состоит из одной вершины, и $K_i = 1$ в случае, когда все узлы являются понятийными.

Экспериментальным путем установлено, что для большинства предметных областей параметр K_i находится в диапазоне от 15% до 30%

Таким образом, третий шаг алгоритма выполняется для тех кластеров, у которых значение $K_i > 0.3$.

C. Третий шаг

Для нахождения центров кластеров предлагается использовать так называемую величину центральности собственных векторов (eigenvector centrality). Для вершины v она есть:

$$x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} x_t,$$

где $M(v)$ - множество вершин смежных с v, а λ - вещественная константа.

Характеристика x_v является мерой воздействия термина на кластер, а именно: если термин является центральным для терминосистемы, то с узлом, задаваемым термином, связано больше понятийных вершин терминологической сети, а значит, для такого узла величина x_v выше.

Относительные оценки x_v вычисляются для каждой вершины в сети, считая, что связь с другим термином с высоким значением данного параметра вносит больший вклад в оценку вершины, чем связь с узлом с низким рейтингом. При этом центром кластера является вершина с наибольшим значением параметра x_v.

Например, исходя из значений параметров x_v, приведенных в таблице I, для вершин кластера, представляющего собой предметную область «Сельское хозяйство», рассматриваемой терминологической сети, можно сделать вывод, что его центром является термин «Сельскохозяйственные культуры», как узел с наибольшим значением параметра центральности. В то же время, понятие «Нелесные земли», скорее всего, относится одновременно к двум предметным областям «Леса» и «Сельское хозяйство».

Таблица I. Наибольшие значения центральности собственных векторов для одного из найденных на шаге 1 кластеров терминологической сети

Термин	Eigenvector centrality
Сельскохозяйственные культуры	1,00
Нелесные земли	0,40
Выращивание растений	0,34
Земли специального назначения	0,33
Технические культуры	0,32
Растениеводство	0,25
Плодовые культуры	0,24

Наконец, в интересах решения проблемы масштабируемости, выделим несколько правил, на основании которых можно определить, что алгоритм произвел ложное слияние двух терминосистем в единый кластер i, а значит следует произвести их разделение:

Правило №1. Если существуют такие понятийные вершины $u, v, w \in V_i$, $deg^+(u) > 1$, $deg^+(v) = 1$, $deg^-(v) = 1$, $deg^+(w) > 1$ и существуют ребра $(u, v) \in E_i$ и $(v, w) \in E_i$ вида «относится-к» (Рис. 2).

Правило №2. Если представить кластер в виде неориентированного графа и в нем найдется хотя бы один мост - ребро удаление которого сделает кластер несвязным.

Рис. 2. Правило №1. Понятийная вершина «Прибрежная растительность» на шаге 1 приводит к объединению терминосистем «Растительность» и «Земная поверхность» в один кластер

IV. Результаты испытания алгоритма

Для тестирования работоспособности описанного подхода использована терминологическая сеть УТП (универсальное терминологическое пространство) [9].

В качестве исходных данных алгоритм использует саму сеть УТП, насчитывающую около 10 тысяч понятийных узлов.

Результатом работы алгоритма является набор терминологических кластеров, который может быть использован в дальнейшем для упрощения ориентации внутри сети. В ходе проверки результата работы алгоритма кластеризации УТП было установлено, что на исходных данных подтверждаются около 93% истинных кластеров.

V. Заключение

В работе рассмотрена задача кластеризации терминологической сети, а также трехэтапная модификация метода Fast unfolding of communities in large networks для поиска сообществ в графе, основанного на максимизации величины модулярности.

На первом этапе работы модифицированного алгоритма производится кластеризация исходного графа, на втором - выделение среди полученных кластеров тех, для которых может иметь место так называемая

проблема масштабируемости, а именно ложного слияния нескольких мелких кластеров в более крупный. Наконец, для выявленных на предыдущем шаге кластеров вычисляются их центры, а также проверяется справедливость правил, свидетельствующих о некорректном слиянии кластеров и потребности в их разделении.

Достоинством рассмотренного алгоритма является его универсальность, то есть возможность воспользоваться им для решения задачи кластеризации применительно к любой терминологической сети независимо от ее предметной области.

В рамках исследования алгоритм был успешно применен к решению задачи кластеризации терминологической сети УТП (универсального терминологического пространства). Результаты проведенного тестирования также приведены в данной работе.

Список литературы

[1] Мальковский М.Г., Соловьев С.Ю. Терминологические сети. // Материалы II Международной научно-технической конференции "Открытые семантические технологии проектирования интеллектуальных систем (OSTIS-2012) Минск: БГУИР, 2012. С.77-82

[2] S. Fortunato. Community detection in graphs. // Physics Reports (2010) Volume: 486, Issue: 3-5.

[3] Чураков А. Н. Анализ социальных сетеи // СоцИс. — 2001. — No 1. — С. 109—12.

[4] M. Girvan, M.E. Newman. Community structure in social and biological networks. // Proceedings of the National Academy of Sciences, 99(12):7821–7826, 2002.

[5] V. D Blondel, Jean-Loup Guillaume, R. Lambiotte, E. Lefebvre. Fast unfolding of communities in large networks. // Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P10008 (12pp)

[6] Соловьев С.Ю. Образные представления терминологической сети. // Сб. Прикладное программное обеспечение. М.: Изд-во МИРЭА, - 2008. стр.55-69

[7] S. Fortunato, M. Barthelemy. Resolution limit in community detection. // Proceedings of the National Academy of Sciences of the United States of America (2007). 104 (1): 36–41.

[8] E. Le Martelot, C. Hankin. Fast Multi-Scale Detection of Relevant Communities in Large Scale Networks. // The Computer Journal, Oxford University Press, 2013. DOI 10.1093/comjnl/bxt002

[9] Мальковский М.Г., Соловьев С.Ю. Исследование родовидовых отношений в терминологических сетях. // Материалы III Международной научно-технической конференции "Открытые семантические технологии проектирования интеллектуальных систем (OSTIS-2013) Минск: БГУИР, 2013. С.147-152

THE STUDY OF TERMINOLOGICAL NETWORK CLUSTERING

Dmitriev K.A.

In this paper I formulate the problem of terminological network clustering and demonstrate one of the possible ways to solve it. In particular, it is proposed a modification of community detection in social networking graphs algorithm, which allows to take into account the peculiar properties of terminological systems and helps to improve the quality of clustering. Moreover, the paper describes a way to estimate the quality of network clustering by calculating a modularity and provides the description of the method that can be used to detect centers of terminological network clusters.

Introduction

In addition to the description of some concepts each term also defines its relationship to other terms within a specific subject area, thereby forming its terminological systems.

With the increasing number of terminological systems added to the terminological network the problem of orientation therein raises caused by the absence of partitioning the network into clusters.

Correctly defined clusters can help the user to navigate quickly through the terminological network.

In terms of machine learning the problem of terminological network clustering is similar to the problem of community detection in social networking graph.

I consider the problem of constructing clusters in terminological networks using methods of social network analysis.

Results and Conclusions

In the paper the problem of clustering terminology network was considered. Moreover, a three-stage modification of the algorithm «Fast unfolding of communities in large networks» for detecting clusters in terminological network, based on maximizing the value of modularity, was proposed.

The advantage of the algorithm is its universality, meaning that it is possible to use it to solve the clustering problem in relation to any terminological network independently from its subject area.

The proposed algorithm has been successfully applied to the problem of clustering terminological network from universal terminology space during the study. The gained results of the testing are also presented in this paper.