ISSN 2658 – 7505
Выпуск №4, 2019
Электронный научный журнал «Вестник молодёжной науки России»

УДК 519.237

ПРИМЕНЕНИЕ МЕТОДОВ КЛАСТЕРНОГО АНАЛИЗА ДЛЯ ИССЛЕДОВАНИЯ СХЕМ ГРУППИРОВАНИЯ ОБЪЕКТОВ

В.И. Ложечка, *loje4kavica@yandex.ru,* ***А.Н. Целых,*** *ant@sfedu.ru*
Инженерно-технологическая академия Южного федерального университета (ИТА ЮФУ),
Таганрог

Кластерный анализ является основным методом для разбиения данных на группы. В данной статье мы рассматриваем применение нескольких алгоритмов данного метода интеллектуального анализа данных. Будут рассматриваться реальные простые графы с неориентированными невзвешенными некратными ребрами. Кластеризация будет рассматриваться, как совокупность математических методов, предназначенных для формирования групп схожих объектов по некому признаку. Данным признаком может являться информация о связях между ними или о расстояниях. Проведенное исследование направлено на идентификацию устойчивых групп, то есть на решение задачи сегментации.

Ключевые слова: кластеризация, сетевой граф, мера центральности, Edge-Betweenness, Label.propagation, Fast Greedy,Walktrap.

THE APPLICATION OF CLUSTER ANALYSIS METHODS FOR RESEARCH OF SCHEMES OF GROUPING OBJECTS

Lozhechka V.I., Tselykh A.N.
Academy for Engineering and Technologies of Southern Federal University,
Taganrog

Cluster analysis is the main method of dividing data into groups. In this paper we consider the application of several algorithms of this method of data mining. We considered simple unweighted undirected graphs without multiple edges. The cluster will be considered as a set of mathematical methods designed to form groups of homogeneous objects on some basis. This function can be information about connections between them or about distances. The aim of the study is to identify stable groups, that is, the solution to the problem of segmentation.

Keywords: clustering, network graph, centrality measure, Edge-Betweenness, Label.propagation, Fast Greedy, Walktrap.

Жизнь человека в современном обществе неразрывно связано с большим количеством данных, в том числе с социальными сетями. Количество пользователей в различных социальных сетях ежедневно растет, а сами социальные сети являются не только «площадкой» для общения, но и доступным политическим, идеологическим и экономическим инструментом. Именно поэтому анализ данных, представляемых в социальных сетях, порождает большой интерес разных исследователей и в настоящее время является одной из важных тем различных научных работ.

Обнаружение сообществ важно для анализа социальных сетей, поскольку с большой степенью вероятности можно утверждать, что узлы в одном сообществе имеют одинаковые свойства. Методы обнаружения сообществ в социальных сетях аналогичны методам и алгоритмам, используемым при разбиении графов на кластеры. Сообщества - это функциональные единицы сети, которые имеют тесные связи внутри, но слабо связаны с внешним миром.

В статье рассмотрены четыре различных алгоритма кластеризации, проведен их анализ и сравнение, представлено теоретическое описание работы алгоритмов и результаты их применения на примере разделения на сообщества графа, созданного на основе данных, взятых из социальной сети «ВКонтакте».

Edge-Betweenness. Алгоритм кластеризации *Edge-Betweenness* [1] представляет собой итеративный процесс, предназначенный для идентификации сплоченных подгрупп. Алгоритм вычисляет значение центральности между всеми ребрами, а затем удаляет ребро или ребра с наибольшим значением данного показателя. Процесс в конечном итоге увеличит количество слабых компонентов, эти компоненты являются связными подгруппами и образуют оптимальное разделение исходных данных на выполняемом шаге. Каждый раз, когда количество компонентов увеличивается, получается новое разделение сети. Процесс продолжается, пока число компонентов не будет меньше указанного пользователем максимума или пока не останется ни одного края.

Алгоритм *Edge-Betweenness* удаляет ребра с наивысшим значением центральности между узлами. Центральность промежуточности основана на понятии кратчайшего пути в графе. Для узла *x* данная мера представляет собой количество кратчайших путей между двумя узлами *i* и *j*, которые проходят через *x*, деленная на общее число кратчайших путей от *i* до *j*.

$$C(x) = \sum_{i \neq j} \frac{\sigma_{ij}(x)}{\sigma_{ij}}$$

Walktrap. Алгоритм *Walktrap* [1] является алгоритмом иерархической кластеризации и принимает в качестве входных данных ориентированные и взвешенные данные. Работа данного алгоритма заключается в прохождении случайным бродягой всех узлов графа по ребрам. Случайным блужданиям обычно разрешается проходить по ребрам,посещать вершины более одного раза или повторять их шаги по только что пройденному ребру.

Теория случайных блужданий основана на том факте, что все короткие случайные блуждания заключены внутри кластера. Если два узла *i* и *j* находятся в одном и том же сообществе, вероятность блуждателя попасть в третий узел *k*, находящийся в том же сообществе, путем случайного обхода не должна сильно отличаться для вершин *i* и *j*. На каждом временном шаге следующий узел выбирается путем случайного выбора соседа текущего узла.

Label.propagation. Label.propagation, или алгоритм распространения меток, (LPA) - это быстрый алгоритм поиска сообществ в графе, обнаруживающий сообщества, используя только сетевую структуру в качестве руководства. Хотя алгоритм не требует предопределенной целевой функции или предварительной информации о сообществах есть возможность назначить предварительные метки, чтобы сузить диапазон создаваемых решений.

Идея, лежащая в основе алгоритма, состоит в том, что одна метка может быстро стать доминирующей в плотно связанной группе узлов, но у нее будут проблемы с пересечением редко связанной области. Метки будут захвачены внутри плотно связанной группы узлов, и те узлы, которые имеют одинаковые метки при завершении работы алгоритма, считаются частью одного сообщества.

ISSN 2658 – 7505
Выпуск №4, 2019
Электронный научный журнал «Вестник молодёжной науки России»

Fast Greedy. Популярным методом обнаружения сообщества является оптимизация модульности. Алгоритм *Fast Greedy* [4] использует для работы базовый жадный подход и основывается на нахождении указанной ранее меры модульности. Работа данного алгоритма начинается с состояние, в котором каждый узел находится в своем сообществе, а алгоритм многократно объединяет некоторые пары сообществ, для формирование более крупных кластеров. Сообщества, отбираемые алгоритмом для слияния, определяются с учетом наибольшего увеличения или наименьшего уменьшения меры модульности. Благодаря описанному агломерационному подходу алгоритм создает набор кластеров с возрастающей степенью детализации.

Значение меры модульности определяется следующим выражением:

$$Q = \sum_{i=1}^{k} \left(e_{ij} - \left(\sum_{j=1}^{k} e_{ij} \right)^2 \right)$$

Экспериментальное исследование. Как было сказано выше, исследование алгоритмов проводилось в среде R. Выбор данной среды программирования не случаен. Основной причиной такого выбора стало наличие широкого выбора пакетов, которые позволяют с легкостью управлять сетевыми данными, а также позволяют выполнять моделирование и визуализацию сетей.

Для проведения анализа был построен социальный граф, который представляет собой связи между друзьями. Для анализа была выбрана социальная сеть «ВКонтакте», так как она является одной из популярных в нашей стране. Все данные для построения графа были взяты с собственно страницы и являются актуальными на момент публикации статьи. Изначально была сформирована бинарная матрица смежности, в которой на главной диагонали все значения были равны 0. На пересечении столбца и строки указывалось, является ли один человек другом второму. Так как получившаяся матрица является громоздкой для наглядности далее приведен лишь фрагмент таблицы.

Таблица 1.

Матрица смежности

	Дукалев	*Semenyuk*	*Рачинский*	*Шевченко*	*Белохвостов*	*Mudruk*
Дукалев	0	1	1	0	1	1
Semenyuk	1	0	1	0	1	1
Рачинский	1	1	0	0	1	0
Шевченко	0	0	0	0	0	0
Белохвостов	1	1	1	0	0	1
Mudruk	1	1	0	0	1	0

Используемая среда программирования позволила с легкостью преобразовать матрицу в сеть, импортируя данные из документа Microsoft Excel. Граф получился небольшой, поэтому можно наглядно увидеть разделение на кластеры. Использование сравнительно не больших сетей для анализа позволяет проследить правильность разделения на сообщества. Получившийся граф имеет 51 вершину и явное разделение части этих вершин на два больших кластера. Для анализа качества кластеризации мы выбрали два важных показателя: модульность, опреде-

ляющая насколько удачно прошел процесс разделения графа, и время, которое понадобилось каждому алгоритму для выявления сообществ. Результаты, получившиеся в результате разбиения можно увидеть в таблице 1.

Таблица 2.

Сравнение параметров кластеризации

	Edge-betweenness	***Walktrap***	***Label. propagation***	***Fast-greedy***
modularity	0.4988114	0.4981129	0.4960421	0.4954049
system.time	0.04	0.05	0.01	0

Разделение сети на кластеры показано на рисунке 1.

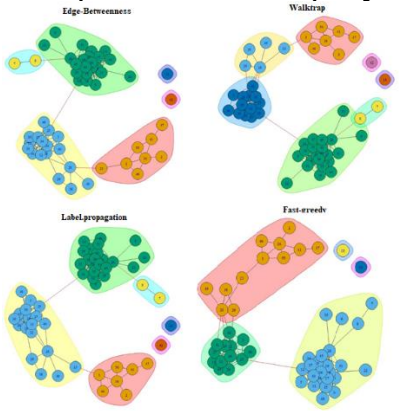

Рис.1. Результат разделения исходной сети на кластеры.

Заключение. Исследуемые в статье алгоритмы очевидно основаны на разных определениях сообщества и поэтому результаты кластеризации отличаются, но все описанные методы направлены на улучшение идентификации сообществ.

В результате проделанного анализа можно явно сказать, что алгоритм *Edge-betweenness*, судя по значению меры модульности, справился с поставленной задачей лучше остальных. Но алгоритм *Walktrap* разделил исходную сеть на большее количество кластеров, что в данном случае было верным решением.

Как описывалось ранее алгоритм *Fast-Greedy* работает значительно быстрее остальных, это можно увидеть и в работе с относительно небольшой сетью, как данная, но не смотря на быстроту алгоритм выделил всего 5 кластеров и показал результаты модульности хуже остальных.

В результате проделанного анализа можно выделить алгоритм *Edge-betweenness*, как лучший среди представленных, и алгоритм *Fast-Greedy*, работающий быстро, но уступающий в качестве разбиения сети.

Список цитируемой литературы

1. Qiaofeng Yang and Stefano Lonardi. A parallel edge-betweenness clustering tool for Protein-Protein Interaction networks // Int. J. Data Mining and Bioinformatics, Vol. 1, No. 3, – 2007. Pages 241-247.
2. Pons P. and Latapy M. Computing communities in large networks using random walks. In Computer and Information Sciences // Journal of Graph Algorithms and Applications, 2005. pages 284–293
3. Xuegang Hu, Wei He, Huizong Li, Jianhan Pan. Role-based Label Propagation Algorithm for Community Detection // School of Computer & Information, Hefei University of Technology, Hefei, China.
4. Ravi Kumar, Benjamin Moseley, Sergei Vassilvitskii // Fast Greedy Algorithms in MapReduce and Streaming. – URL: http://cseweb.ucsd.edu/~avattani/papers/mrgreedy.pdf