Министерство образования и науки Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

АЛГОРИТМЫ ПОИСКА СООБЩЕСТВ В НЕОРИЕНТИРОВАННЫХ ГРАФАХ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента 2 курса 273 группы
направления 01.04.02 Прикладная математика и информатика
факультета компьютерных наук и информационных технологий
Ионкина Михаила Сергеевича

Научный руководитель:
зав. кафедрой Информатики
и программирования, к.ф.-м.н. _____ Огнева М. В.

подпись, дата

Зав. кафедрой:
к.ф.-м.н. _____ Огнева М. В.

подпись, дата

Саратов 2017

ВВЕДЕНИЕ

Актуальность темы

В настоящий момент наблюдается значительный рост объема данных. Ежедневно пользователи Интернета посещают множество сайтов, отмечают на географических картах свое местоположение, обмениваются письмами, мультимедиа файлами, документами, общаются в социальных сетях и т.д. Возникают новые задачи, связанные с обработкой и анализом больших объемов данных, а, следовательно, появляются новые инструменты и технологии для решения таких задач.

Одной из основных задач анализа данных является задача кластеризации (или кластерный анализ) - выделение сообществ (кластеров) разных объектов: пользователей, сайтов, продуктов интернет-магазинов и так далее. Например, с помощью нее можно находить группы пользователей с похожими предпочтениями, что в свою очередь помогает определить, какая информация будет для них наиболее интересна. Данный подход применяется в ходе маркетинговых исследований [1]. Методы кластеризации можно также использовать для сегментации изображений (что необходимо для так называемой технологии компьютерного зрения), для распознавания образов и рукописного текста, для извлечения информации и для многого другого [1, 2, 3].

В ходе анализа доступной литературы и публикаций был сделан вывод, что, несмотря на актуальность задачи кластеризации, она до сих пор не решена окончательно [4]. В представленных источниках описано множество алгоритмов для решения данной проблемы [1, 5, 6, 7]. И среди них нет универсального – каждый имеет свои ограничения, преимущества и недостатки. Так, в книгах [3, 6] и на Интернет ресурсах [8, 9] утверждается, что некоторые алгоритмы требуют априорных знаний о будущих выходных данных (например, о числе получаемых кластеров). В статье [10] авторы пытаются решить проблему плохой масштабируемости этих алгоритмов на большие объемы данных. В статье [11] авторы предлагают свое решение

проблемы нелинейного времени выполнения, разработав иерархический алгоритм, имеющий логарифмическую сложность. В статье [12] авторы приводят алгоритм, который позволяет находить кластеры в графах основываясь не только на связи между вершинами, но и учитывая атрибуты этих вершин. А в источниках [1, 13] приводят рекомендации по применению алгоритмов кластеризации в зависимости от структуры или типа входных данных (например, направленные графы, изображения, тексты).

Также было обнаружено, что на данный момент не существует наилучшего формального критерия оценки качества кластеризации данных. Известен целый ряд эвристических критериев, и все они могут давать разные результаты [1, 14]. Следовательно, для определения качества кластеризации требуется эксперт предметной области, который бы мог оценить осмысленность выделения кластеров. Сам результат кластеризации существенно зависит от метрики (меры близости между двумя объектами), выбор которой, как правило, также субъективен и определяется исследователем в зависимости от количества атрибутов, типа данных и их структуры [1, 15].

В связи с этим возникла потребность в сравнении существующих алгоритмов, в выявлении их преимуществ и недостатков, что позволит выявить схожие принципы их работы и, возможно, в дальнейшем приведет к созданию универсального алгоритма.

Цель магистерской работы – выполнение сравнительного анализа алгоритмов поиска сообществ в графах.

Поставленная цель определила **следующие задачи**:

1. Описать постановку задачи машинного обучения и, в частности, кластеризации;
2. Изучить наиболее популярные алгоритмы поиска сообществ в графах, дать их краткое описание, а также найти существующие реализации или реализовать самостоятельно;

3. Рассмотреть методы поиска расстояний между объектами применимые для задачи кластеризации;

4. Описать способы оценки качества кластеризации;

5. Попытаться улучшить алгоритм поиска сообществ для графов с взвешенными вершинами;

6. Подобрать наглядные примеры и данные с ground-truth сообществами для анализа алгоритмов;

7. Выполнить анализ результатов работы алгоритмов «вручную», то есть, не используя каких-либо автоматических средств проверки качества;

8. Выполнить анализ результатов работы алгоритмов, используя наиболее популярные метрики и функционалы качества.

Структура и объём работы. Магистерская работа состоит из введения, 4 разделов, заключения, списка использованных источников и 12 приложений. Общий объем работы – 95 страниц, из них 64 страницы – основное содержание, включая 28 рисунков и 7 таблиц, список использованных источников информации – 39 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Интеллектуальный анализ данных и машинное обучение» посвящен обзору основных понятий, связанных с интеллектуальным анализом данных и машинном обучением. Рассмотрено понятие «большие данные» и некоторые техники анализа, применимые к ним. Приведена краткая история возникновения термина «data mining», дана общая постановка задачи data mining и классификация таких задач. Приведено формальное определение задачи обучения по прецедентам (обучение с учителем) и задачи кластеризации (обучение без учителя).

Второй раздел «Поиск сообществ в неориентированных графах» посвящен описанию одной из задач машинного обучения, а именно, - поиску сообществ в неориентированных графах. В начале раздела описывается понятие «сообщество в графе» и рассматриваются проблемы, которые

возникают из-за того, что не существует формального определения понятия «сообщество в графе». Затрагивается проблема оценки результатов работы алгоритмов. Далее рассматриваются самые популярные на данный момент алгоритмы для поиска сообществ: Infomap, Walktrap, Label Propagation, Fastgreedy, Edge Betweeness, Louvain, Smart Local Moving. Делается вывод о том, что неопределенность постановки задачи поиска сообществ в графах приводит к разнообразию подходов к решению поставленной задачи и к разнообразию методов оценивания алгоритмов.

Третий раздел «Метрики и оценка качества кластеризации» посвящен более подробному рассмотрению проблемы оценки качества кластеризации. Здесь описываются некоторые стандартные метрики, применяемые в алгоритмах кластеризации, дается определение понятий «функция потерь» и «функционал качества», описывается «модулярность» - наиболее популярный функционал качества для задачи поиска сообществ в графах, а также рассматривается «нормализованная взаимная информация» (NMI) - популярный в последнее время способ сравнения разных разбиений. Приводятся преимущества и недостатки модулярности и нормализованной взаимной информации.

Четвертый раздел «Практическая часть» посвящен практическому применению рассмотренных алгоритмов, оценке их характеристик и проведению сравнительного анализа.

Один из самых известных на данный момент алгоритмов поиска сообществ Louvain использует в своей основе модулярность. Обычно алгоритмы, использующие модулярность не замечают мелкие (относительно всего графа) сообщества и объединяют их в одно сообщество. Эта проблема получила название resolution limit. Существует несколько работ, посвященных исследованию этой проблемы и способам ее решения (например, в [16, 17, 18, 19]). В ходе экспериментов выяснилось, что помимо обозначенной выше проблемы, на небольших объемах данных алгоритм Louvain делит их на слишком мелкие сообщества. И можно сказать, что эта

проблема является прямо противоположной проблеме resolution limit. Для решения этой проблемы было предложено учитывать веса вершин графа при подсчете модулярности во время работы алгоритма Louvain. Был проведен эксперимент, в результате которого, расставив веса вершин у графа, удалось сократить количество кластеров в получающемся разбиении.

Далее было проведено сравнение алгоритмов Infomap, Walktrap, Label Propagation, Fastgreedy, Edge Betweeness, Louvain на «искусственных» данных. Первый набор данных представлял собой граф, состоящий из двух не связанных между собой циклов, а второй набор – известный граф «клуб карате Zachary». Сравнение производилось с помощью модулярности и «вручную» (то есть оба набора данных были заранее разбиты на сообщества, которые в дальнейшем сравнивались с разбиениями, получающимися в результате работы алгоритмов). В результате выполнения двух тестов видно, что результаты работы алгоритмов зависят от структуры графа. Например, алгоритм Fastgreedy показал лучший результат при анализе графа «клуб карате», но при этом плохо справился с анализом графа, состоящего из двух циклов, не соединенных между собой. Но, например, результаты таких алгоритмов как Infomap и Louvain были приемлемого качества в этих двух тестах. Также видно, что решение, лучшее с нашей точки зрения, не всегда совпадает решением, лучшим с точки зрения функционала качества. Отсюда следует, что необходимо учитывать принцип работы функционала. Например, сети с высокой модулярностью имеют плотные связи между вершинами внутри сообществ, но редкие связи между вершинами из разных сообществ. Но ведь не всегда плотные связи внутри одного сообщества означают, что его вершины действительно должны ему принадлежать.

В ходе анализа доступной литературы и публикаций было обнаружено, что алгоритм Smart Local Moving (SLM) обладает преимуществом по сравнению с известным алгоритмом Louvain: он позволяет находить разбиение графа с большей модулярностью, чем Louvain. Поэтому было решено реализовать данный алгоритм на языке программирования Python,

чтобы была возможность выполнять сравнительные тесты с другими уже реализованными алгоритмами из библиотеки igraph в одной среде. Это позволит не учитывать в ходе анализа, например, различные механизмы сборки мусора языков программирования, время, затрачиваемое на компиляцию или интерпретацию исходного кода, внутреннюю реализацию базовых структур данных языков и т.д. В ходе реализации использовалось описание алгоритма SLM и его псевдокод (приложение А) из статьи [20].

После анализа работы всех перечисленных выше алгоритмов на «искусственных» примерах следующим очевидным шагом является выполнение этого же анализа, но уже на «реальных» данных. Также было решено сравнить получившиеся сообщества каждого отдельного алгоритма с ground-truth сообществами, то есть с уже известными сообществами этой сети, сформированными по какому-либо признаку (например, группа в социальной сети, в которой состоят поклонники какого-нибудь артиста).

В качестве таких данных были взяты общедоступные данные трех сайтов (You Tube, Amazon и Live Journal) [21], которые представляют собой большие разреженные графы. Все данные «обезличены» и представляют собой просто набор идентификаторов. Было проведено сравнение результатов работы алгоритмов по их времени выполнения и по значению модулярности отдельно для каждого набора данных. Время выполнения всех алгоритмов совпало с теоретическими оценками их времени выполнения. С помощью метрики «нормализованная взаимная информация» было проведено сравнение получившихся разбиений с ground-truth сообществами. Так как у некоторых алгоритмов мы получили примерно одинаковые значения модулярности и NMI, то необходимо было выяснить получаются ли разбиения одинаковыми или это другие разбиения, но с похожей модулярностью. Для этого было проведено сравнение разбиений алгоритмов между собой с помощью метрики NMI. Это показало, например, что алгоритмы Louvain и SLM, а также Walktrap и Label Propagation находят наиболее похожие разбиения, не зависимо от набора данных.

Так как предыдущий анализ выполнялся на «обезличенных» данных, то не совсем понятно какой смысл несут в себе получившиеся разбиения. Поэтому для дальнейшего анализа был решено использовать данные социальной сети «ВКонтакте» (далее по тексту – набор данных Vk). Для этого был написан модуль на языке Python для загрузки и предварительной обработки информации о пользователях этой социальной сети.

Рассматриваемый граф является моделью отношения “является другом” среди пользователей социальной сети. Был взят один пользователь социальной сети и список всех его друзей, и построен неориентированный граф следующим образом: вершинами являются друзья этого пользователя, и между двумя вершинами проводится ребро, если эти два человека являются друг другу друзьями. Далее получившийся граф был вручную поделен на ground-truth сообщества. Был проведен анализ результатов работы алгоритмов на этом наборе данных с помощью модулярности и NMI. Было проведено сравнение получившихся разбиений с ground-truth сообществами и между собой. В данном примере алгоритмы Lovain и SLM смогли достаточно точно определить социальные группы среди друзей пользователя, а наиболее похожие получились разбиения у алгоритмов Walktrap, Fastgreedy, Edge Betweenness и Infomap и их разбиения довольно сильно отличаются от ground-truth сообществ.

Был сделан вывод, что алгоритмы способны, основываясь только на знании о структуре графа, разбивать его на сообщества, которые несут в себе ценную информацию, которая не была доступна ни исследователю, ни самим алгоритмам. А в этом и заключается основная идея data mining.

ЗАКЛЮЧЕНИЕ

В работе проведен анализ наиболее популярных алгоритмов поиска сообществ в графах, изучены их особенности выполнения и свойства, дано краткое описание их работы, подобраны необходимые библиотеки для языка программирования Python.

Было приведено сравнение результатов кластеризации алгоритма Louvain с «ручной» обработкой данных. Также в работе предложен вариант решения проблемы, противоположной проблеме resolution limit – нахождение мелких относительно всего графа сообществ. Для это было предложено учитывать веса вершин графа, что помогло подсчитывать модулярность более точно и, таким образом, производить более качественное разделение графа на сообщества (для кластеризации использовался алгоритм Louvain).

Проанализирована работа выбранных алгоритмов на двух тестах (первый – граф, состоящий из двух циклов, второй – один из самых известных наборов данных «клуб карате»). Анализ качества работы алгоритмов осуществлялся «вручную», то есть без использования функционалов качества.

На языке Python был реализован алгоритм Smart Local Moving, который является улучшением алгоритма Louvain в максимизации модулярности.

Был проведен сравнительный анализ выбранных алгоритмов на трех наборах данных с ground-truth сообществами, содержащими несколько сотен тысяч вершин и ребер, представляющие собой большие разреженные графы. Оценивались такие характеристики как время выполнения алгоритмов, показатели модулярности и нормализованной взаимной информации (NMI). Так как данные этих трех наборов были «обезличены», то было решено провести подобный анализ на данных социальной сети «ВКонтакте». Для этого был написан модуль на языке Python для загрузки информации о пользователях сайта. Все конкретные выводы и результаты относительно работы алгоритмов представлены в соответствующих главах данной работы. Обобщив, можно сказать, что каждый алгоритм имеет свои преимущества и недостатки. Например, по времени выполнения и максимизации значения модулярности явными лидерами являются алгоритмы Louvain и Smart Local Moving. При этом они не всегда хорошо находят ground-truth сообщества. С этой задачей в двух тестах хорошо справился алгоритм Fastgreedy, но не смог найти «правильную» структуру для данных социальной сети «ВКонтакте».

В ходе исследования был сделан вывод, что, не смотря на актуальность задачи поиска сообществ, она до сих пор не решена окончательно. И среди представленных решений нет универсального – каждый имеет свои ограничения, преимущества и недостатки.

По тематике магистерской работы были представлены доклады:

1. Ионкин М.С., Огнева М.В. Алгоритм Louvain для поиска сообществ в графах с взвешенными вершинами // Актуальные проблемы автоматизации и управления в технических и организационных системах (АПАУ-2016): сб. тр. междунар. науч. конф. / под ред. М. Ф. Степанова. Саратов: «Амиринт», 2016. – С. 90-96.
2. Ионкин М.С., Огнева М.В. Основы интеллектуального анализа данных для школьников // Информационные технологии в образовании: Материалы VIII Всерос. научно-практ. конф. – Саратов: ООО «Издательский центр «Наука»», 2016. – С. 340-344.
3. I Международная научно-практическая конференция «Электронные образовательные технологии – пространство неограниченных возможностей» 16-17 марта 2017 г., г. Новосибирск, «Непрерывная подготовка специалистов в области анализа данных»

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Aggarwal, C. C. Data clustering. Algorithms and applications / Charu C. Aggarwal, Chandan K. Reddy. N.-Y.: Chapman and Hall/CRC, 2014. — 652 p.
2. Jain, A. K., Murty M. N., Flynn, P. J. Data clustering: a review // Acm computing surveys. 1999. vol. 31, № 3. pp. 264-323.
3. Сегаран, Т. Программируем коллективный разум. / Т. Сегаран. – пер. с англ. – Спб: Символ-плюс, 2008. – 368 с.
4. Федоренко, Ю. С. Кластеризация данных на основе нейронного газа и марковских алгоритмов // Молодежный научно-технический вестник. 2014. № 8.

5. Newman, M. E. J. Detecting community structure in networks // The European Physical Journal B - Condensed Matter and Complex Systems. 2004. Volume 38, issue 2. pp. 321–330.
6. Leskovec, J. Mining of massive datasets / Jure Leskovec, Anand Rajaraman, Jeff Ullman; 2nd edition. Cambridge University Press, 2014. – 511 p.
7. Fortunato, S. Community detection in graphs // Physics Reports. 2010. № 486(3). pp. 75-174.
8. Информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных [Электронный ресурс] : [сайт]. URL: www.machinelearning.ru (дата обращения 12.12.2015). Загл. с экрана.
9. Чубукова, И. А. Курс лекций «data mining» // Интернет-университет информационных технологий [Электронный ресурс] : [сайт]. URL: www.intuit.ru/department/database/datamining (дата обращения 02.02.2015). Загл. с экрана.
10. Arnau Prat-Pérez, David Dominguez-Sal , Josep-Lluis Larriba-Pey. High quality, scalable and parallel community detection for large real graphs // Proceedings of the 23rd international conference on World Wide Web. 2014. pp. 225-236.
11. Clauset, A., Newman, M. E. J., Moore, C. Finding community structure in very large networks // Physical Review. 2004. № 70(066111).
12. Jaewon Yang, Julian Mcauley, Jure Leskovec. Community detection in networks with node attributes // IEEE 13th International Conference on Data Mining. 2013. pp. 1151-1156.
13. Witten I. H., Frank E., Hall M. A. Data mining: practical machine learning tools and techniques / Ian H. Witten, Eibe Frank, Mark A. Hall; 3rd edition - San Francisco: Morgan Kaufmann Publishers Inc., - 2011. – pp. 665.
14. Fragkiskos d. Malliaros, Michalis Vazirgiannis. Clustering and community detection in directed networks: a survey // Physics Reports. 2013. Volume 533, issue 4. pp. 95-142.

15. Tan, P.-N., Steinbach, M., Vipin, K. Introduction to Data Mining / Pang-Ning Tan, Michael Steinbach, Vipin Kumar, - 1st edition - Boston, MA, USA : Addison-Wesley Longman Publishing Co., Inc. – 2005. - pp. 725.
16. Santo Fortunato, Marc Barthélemy. Resolution limit in community detection // Proceedings of the National Academy of Sciences. 2007. № 104. pp 36-41.
17. Kumpula, J. M., Saramaki, J., Kaski, K., Kertesz, J. Limited resolution and multiresolution methods in complex network community detection // Fluctuation Noise Letters. 2007. № 7 (209).
18. Ronhovde, P., Nussinov, Z. Local resolution-limit-free potts model for community detection // Physical Review. 2010. № 81.
19. Traag, V. A., Dooren, P. V., Nesterov, Y. Narrow scope for resolution-limit-free community detection. 2011. Physical Review. № 84.
20. Waltman, L., Van Eck, N.J. A smart local moving algorithm for large-scale modularity-based community detection // The European Physical Journal B. 2013. Volume 86 (11).
21. Jure Leskovec. Stanford Large Network Dataset Collection // Stanford Network Analysis Project [Электронный ресурс] : [сайт]. URL: https://snap.stanford.edu/data (дата обращения 25.04.2017). Загл. с экрана.