Fields of Mathematics classified by Wikipedia

Fall 2021 CSCI 4022 Final Project

Euijin Hong

December 16, 2021

1 Introduction

The branches of a field of study in these days are so diverse, specified and inter-related with many area of studies. So, it is hard to determine if a topic from a discipline belongs to any of its sub field. Wikipedia is the biggest online encyclopedia which contains information about almost anything that can be documented with experts' level of depth. Moreover, it consists of hyperlinked web pages, any two topics closely related to each other have a link from one to another of vice versa. Web pages of Wikipedia form a massive network itself. So, if you can get an information of the connections between web pages, we can analyze the topics in Wikipedia using techniques of classifying communities in networks. In this project, using links between popular mathematics topics in Wikipedia pages, we try to classify the sub fields of mathematics by its topics.

2 Data

The data of the project can be obtained from Machine Learning Repository website of University of California Irvine. There, on April 20, 2021, Wikipedia Math Essentials data set was posted [Roz+21]. The data contains information of nodes and edges, where each node corresponds to a Wikipedia's popular website in the branch of mathematics and any two nodes connected by links have a corresponding edge. The edges don't contain the information of in-link and out-link, so the data only gives undirected graph of nodes and edges. It also contains weights of each edge based on the number of links between two nodes. In addition to that, the data has number of visits of each website per day from the March 2019 to 2021.

The data was given in the type of dictionary which contains a total of 735 keys. This project only use the first 3 keys, which are 'edges', 'weights' and 'node_ids'. The key of 'edges' is a list of tuples of nodes, 'weights' is the list of integers with the same length as 'edges', and 'node_ids' is the dictionary of topic and node numbers. Some of random examples of 'node_ids' are give below.

{'Continuous function': 63,
 'Multiset': 696,
 'Tree (graph theory)': 474,
 'Eigenvalues and eigenvectors': 796,
 'Isomorphism': 412,
 'Elliptic partial differential equation': 917}

3 Methods

In this project, we want to classify topics of mathematics into its corresponding sub fields. For example, we want to classify the topic 'module' under 'Algebra', 'Hyperbolic geometry' under 'Geometry' and 'Liouville's theorem' under 'Analysis'. To this, we will implement the graph partitioning algorithm. However, it is obviously possible for a topic to belong to multiple sub areas of mathematics. For example, 'Homological algebra' can belong both to 'Algebra' and 'Topoloby'. To this, I also implement the BigCLAM algorithm to find the overlapping communities of topics.

3.1 Graph Partition

Tow different methods used for multi class-classifications. One for using Fiedler vector and cut where the largest jump happens and keep cutting the next largest jump position. Another method is to use multiple eigenvectors of the Laplacian matrix and divide the nodes based on whether the corresponding position of the eigenvector has positive or negative values.

Graph Partition algorithm is implemented for hard clustering. The algorithm runs as follow:

1. Construct the adjacency matrix A, degree matrix D and Laplacian matrix L.
2. Choose the eigenvectors corresponding to the 2nd to n-th smallest eigenvalues.
3. (a) Method 1. Only use the Fiedler vector and cut some of the largest jumps. For each cut one more cluster is created.
   (b) Method 2. Use 2nd to n-th eigenvectors and classify nodes based on whether the corresponding position in eigenvector has positive values or not. For each cut, the number of clusters can be double theoretically.

3.2 BigCLAM

The algorithm BigCLAM computes the maximum likelihood of affiliation of all nodes to all communities. The basic idea is, one's neighbourhood has higher probability of beinng in one community, then that nodes will get higher score of being in the same community. The gradient ascent method is implemented to gradually update community association scores for each node until the scores will converge within certain error.

4 Results

4.1 Graph Partition

Tow separate methods are implemented. The first one is to cut the graph where the maximal gaps of values of Fiedler vector happens.

4.1.1 Multiple cut by maximal gaps of the Fiedler vector

- 2 clusters obtained by 1 cut

Cluster 1	Kolmogorov structure function
Cluster 2	All the rest nodes

- 4 clusters obtained by 3 cuts

Cluster 1	Kolmogorov structure function
Cluster 2	Kolmogorov complexity
Cluster 3	Cache replacement policies
Cluster 4	All the rest nodes

- 8 clusters obtained by 7 cuts

Cluster 1	Kolmogorov structure function
Cluster 2	Kolmogorov complexity
Cluster 3	Cache replacement policies
Cluster 4	Blum axioms
Cluster 5	Data compression
Cluster 6	Pigeonhole principle, Entropy (information theory) Model of computation, Gap theorem
Cluster 7	Universal Turing machine, Gödel numbering, Telephone number Shellsort, Blum's speedup theorem
Cluster 8	All the rest nodes

Here, we can see that clustring by cutting the maximal jump initially only separates clusters of single nodes. This phenomenon can be observed if we plot the values of Fiedler vector. Figure 1 shows how the values are distributed. The top left one shows all values and one below shows rest of values after cutting the largest jump and it continues to the next column. Until we cut the initial 5 largest jump, only one node is separated from the main cluster.

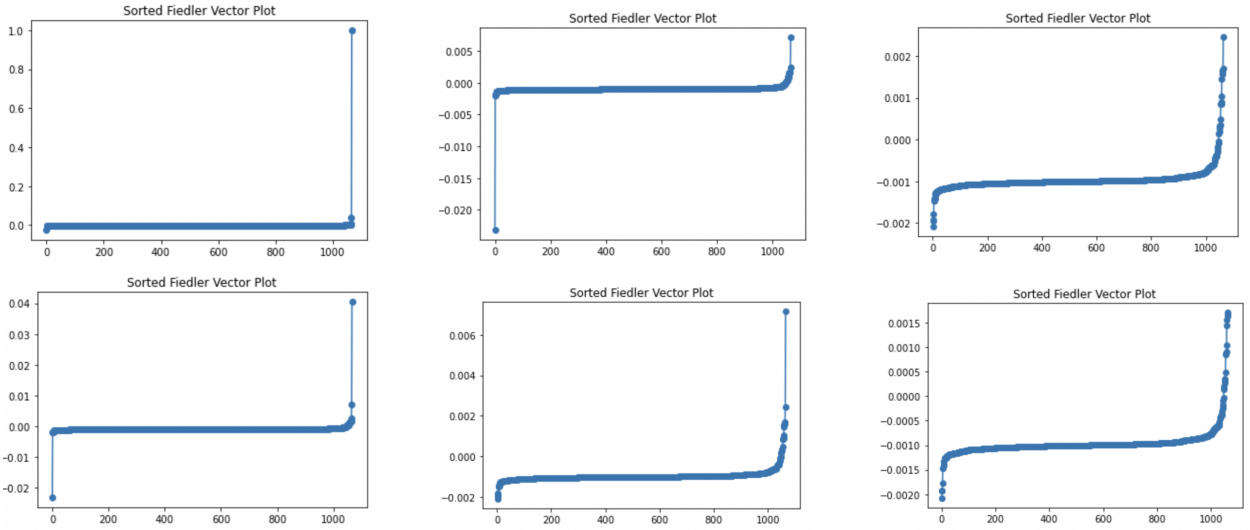

Figure 1: Values of Fiedler Vector

4.1.2 Optimal cut by multiple eigenvectors

The second method of multiple Graph Partition is implemented by using multiple eigenvectors. First, cluster the nodes which have positive value at its corresponding position of the Fiedler vector. Then repeat the same process with the next eigenvector, whose eigenvalue is the smallest after the Fiedler vector. This method gives the follwogin results.

- Optimal cut by the Fiedler vector.

Cluster 1	Gödel's incompleteness theorems, Fast Fourier transform, Turing machine, Data compression, Entropy (information theory), Model of computation, Data structure, Pigeonhole principle, Halting problem, Mandelbrot set, Kolmogorov complexity, Blum axioms, Blum's speedup theorem, Universal Turing machine, Gödel numbering, Telephone number, Shellsort, UTM theorem, Kolmogorov structure function, Gap theorem
Cluster 2	All the rest nodes

- Optimal cut by 4 eigenvectors The following table shows some clustering of optimal cuts by four eigenvectors. Theoretically there can be a total of $2^4 = 16$ clusters, but here, 11 clusters are formed.

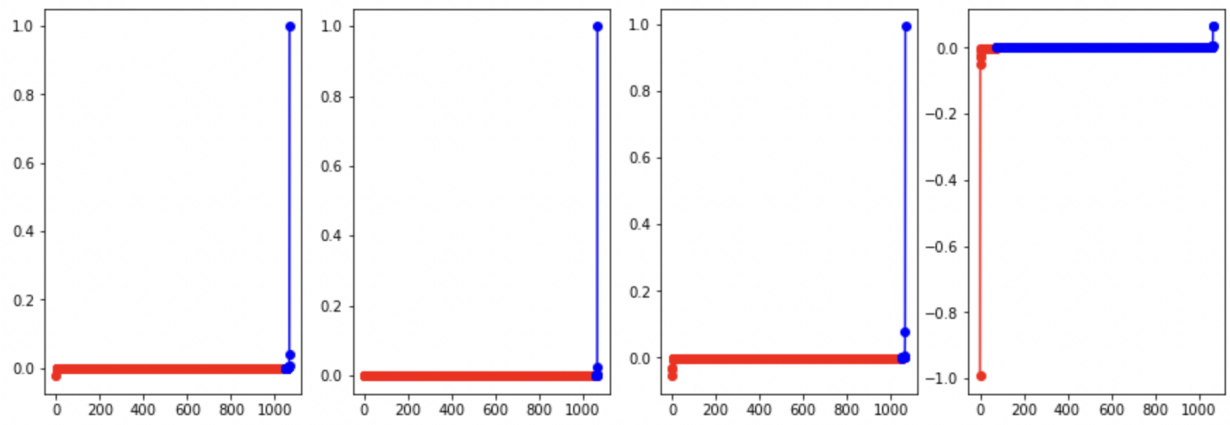

Figure 2: Sorted values of Fiedler vector and the following three eigenvectors

Cluster 1	Theoretical computer science, Information theory, Number, Fourier transform, Markov chain, Coding theory, Numeral system, Tally marks, Arabic numerals, Decimal separator
Cluster 2	Statistics,Design of experiments,Observational study,Statistical hypothesis testing,Fourier analysis,Probability distribution,Random variable,Bayes' theorem,Normal distribution,Probability density function,Central limit theorem
and more	

This seems better than cut by maximal jumps of the Fiedler vector. However, there are still a big cluster of nodes remained and the number of nodes are imbalanced by the cuts also. This also can be visible by checking the values of the eigenvectors starting from the Fiedler vector. The Figure 2 shows this where negative values are plotted by red dots.

In practice, the multiple cut of graph is a mixture of the two methods, cut at the largest jump of an eigenvector and then move to the next eigenvector and repeat the first process. In this case, the clustering will be the intersection of the results from the first and the second process.

4.2 BigCLAM

It is noted in the code that without initialization, all nodes have the same community affiliation scores for all communities. So, the result is presented with distinct initialization. Also, we first present the result of hard clustering, that is, any node is assigned to whichever community with the highest association score.

4.2.1 Hard Clustering by BigCLAM

- Initialize 'Algebra', 'Geometry' and 'Mathematical Analysis'

Algebra	Geometry	Analysis
467	245	356
Galois Theory	Parabola	Probability Theory
Representation Theory	Triangle	Hilbert Space
Higher-order logic	Fermat's Last Theorem	Algebraic function
Topological Space	Algebraic Number field	
2	3	Pi
7	Mathematician	E

Here, the numbers below 'Algebra' or so indicates the number of nodes assigned to each cluster. By initialization they are classified evenly compare to any of the previous methods. The two lines below indicates well-classified nodes, which we can easily expect to be classified in that cluster. The next two lines are examples of something not obvious or awkward. For example, 'Higher-order logic' may be expected to classified in 'Algebra' among the three initialization but in general, we may need more communities to classify Mathematical Logic or Set theory independently. 'Topological Space', 'Algebraic Number field' or 'Algebraic function' better be classified under different category. The next two lines are neutral topics. It doesn't belong to any specific branch of mathematics since they are so general to be restricted under any of specific subfield of mathematics.

- Add 'Topology' and 'Set theory' to the initialization

Algebra, Geometry, Analysis	Topology	Set Theory
	77	319
⋯	Top'l space	Cardinal Number
⋯	Homotopy	Zorn's Lemma
⋯	Alg Num Fld	Depth-first search
Yoneda Lem	Hom'l Alg	Hom functor
⋯	Functor	Forgetful functor

As more communities added with proper initialization, some topics are classified correctly. However, still some closely related topics are spread over different clusters. For example, the topics of bottom two rows are all belong to Category theory, but they are all spread out.

- Further add 'Applied Mathematics' and 'Statistics' to the initialization.

Previous 5	Applied Math	Statistics
	101	125
⋯	algorithms	Bayes' theorem
⋯	LU Decomposition	Monty Hall problem
·	Queue, Heap, Quicksort	Logistic regression
⋯	Pure Math	Dirac delta function
⋯	PDE, ODE	Gamma function
⋯	Functional analysis	Logarithm

One interesting result of this classification is that 'Pure Mathematics' is classified under 'Applied Mathematics'. This wasn't expected, can be easily explained. Any field of study may show its characteristic by contrasting with its opposite component. So, this kind of peculiar connection may happen.

4.2.2 Soft Clustering by BigCLAM

Some result of soft clustering is presented here. The communities with the top two affiliation scores are listed.

Fermat's Last Theorem	Algebra, Geometry
Algebraic Geometry	Geometry, Topology
Gödel's incompleteness theorems	Algebra, Set Theory
Homological Algebra	Algebra, Topology
Functor	Topology, Set Theory
PDE	Analysis, Topology

5 Conclusion and Further Development

The method above can be applied with other branches of academic discipline. Or even further, it can be applied to any field or all links of Wikipedia and we may draw a map of knowledge, which is not systematically constructed but inadvertently built by experts of each topics.

Some interesting features we noticed are

- Graph Partition may not give seems to be 'optimal' cut possibly because of pathological isolated islands of nodes.
- Exactly opposite topics can be classified under the same branch because the nature of being opposite helps each other more definite.

I also want to suggest possible further project based on this.

- There are hierarchies between topics. For example, we may order 'Mathematics' – 'Algebra' – 'Number Theory' – 'Fermat's Last Theorem' from general to specific one. If we can rank each node by its hierarchies, we can create a learning path from the beginning to deeper theory of a specified topic.
- Extends the nodes and edges to other field of study and then find interdisciplinary fields of study. We can also figure out what knowledge is needed from each field of study. For example, what CS, math, or stat knowledge are required to take Data Science class. This example is already well known, I guess, but we can further search it for other topics/areas also.

References

[Roz+21] Benedek Rozemberczki et al. *PyTorch Geometric Temporal: Spatiotemporal Signal Processing with Neural Machine Learning Models*. 2021. eprint: [WebLink].