

# Analysis of Study Patterns through Wikipedia

BASIL KHWAJA<sup>1</sup>

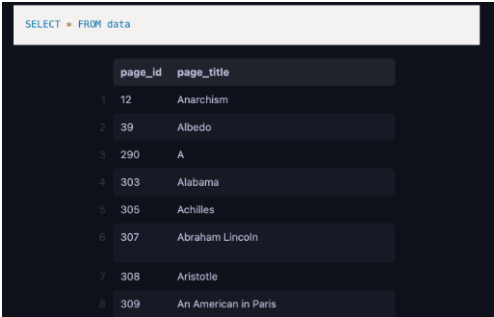
<sup>1</sup>*Purdue University*

## 1. INTRODUCTION

In todays fast-paced academic environment, students often struggle to find effective study strategies and like-minded peers to collaborate with. This project explores the development of a platform that enhances the study experience by intelligently recommending study resources and connecting students who are learning similar topics. By analyzing user behavior, such as course engagement, study patterns, and topic preferences, the system builds a dynamic knowledge graph linking students with relevant materials and peers. We use techniques from graph theory, recommendation systems, and user modeling to show that this platform aims to helps students increase productivity and find like-minded peers. Ultimately, this work seeks to reduce academic isolation, improve resource discovery, and empower students to study more effectively through community-driven support.

## 2. CONTEXT

One of the largest and most accessible sources of information on the Internet today is Wikipedia. While some view it with skepticism in academic settings, often raising concerns about reliability or the potential for misinformation, it still holds significant value for students who are just beginning to explore a topic. Rather than serving as a final authority, Wikipedia works best as a starting point, offering a general overview before diving into more specialized or peer-reviewed material.



SELECT \* FROM data

	page_id	page_title
1	12	Anarchism
2	39	Albedo
3	290	A
4	303	Alabama
5	305	Achilles
6	307	Abraham Lincoln
7	308	Aristotle
8	309	An American in Paris

**Figure 1.** Dataset for correlation of id to wiki page names



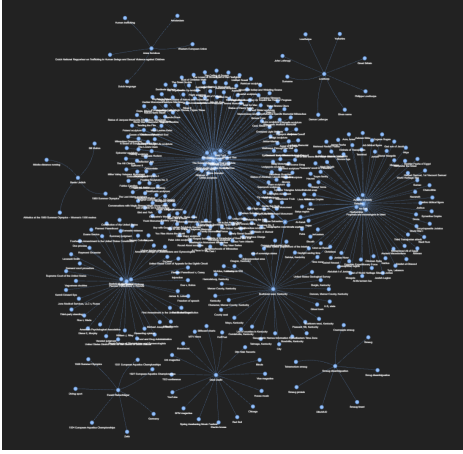
	pl_from	pl_to
1	12	308
2	12	339
3	12	988
4	12	1023
5	12	1167
6	12	1193
7	12	1216
8	12	1814
9	12	2023

**Figure 2.** The links between different wiki pages

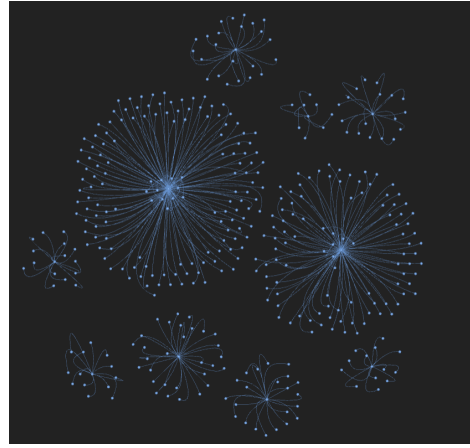
What makes Wikipedia particularly useful is the way its content is interconnected. Each article is not isolated, but linked to many others through internal hyperlinks. This allows users to easily explore related topics and build a

broader understanding by simply following these links. When you begin to think of Wikipedia in this way, it becomes more than just an online encyclopedia. It starts to resemble a dynamic and evolving knowledge graph, where each page acts as a node and each hyperlink represents a connection between concepts. By capturing this structure, we can build a dataset that includes both the content of each Wikipedia page and the relationships it shares with other pages. This can be incredibly powerful for learners, as it allows them not just to read about a single topic, but to see how it connects to a larger web of information. It can also help surface related ideas that they might not have discovered otherwise, encouraging deeper and more connected learning.

### 3. METHODOLOGY



**Figure 3.** Force Atlas 2 Based Graph



**Figure 4.** Hierarchical Repulsion Graph

To effectively build and visualize a graph of Wikipedia pages and their connections, we can take advantage of the combined strengths of NetworkX and PyVis. This makes it ideal for organizing and understanding how different Wikipedia pages relate to one another. Once we have the graph built using NetworkX, we can use PyVis to turn that structure into an interactive visualization. PyVis works well with graphs created in NetworkX, making it easy to take the data and render it in a browser. By using NetworkX to handle the data and PyVis to handle the presentation, we can create a tool that is both powerful and accessible. This combination is especially valuable in educational settings, where students can not only build and analyze networks but also explore them visually. For example, a student might start with a single Wikipedia page they are interested in and use the visualization to discover related topics, identify clusters of information, or trace connections between ideas. This approach supports a more interactive and intuitive learning experience that goes beyond reading static text.

Some limitations that do exist with the visualization of these graphs is they are computationally costly to create and render. While Figure 3 and Fig 4 show a lot of information of a singular user, the depth of the information we are obtained has a max distance of 1. If we are able to obtain higher-end GPUs like the H100 from nvidia we perform high level calculations on the graph.

### 4. EXPERIMENTATION

In order to see the benefits of converting a knowledge graph structure from the dataset on Wikipedia, we simulated a platform where users can interact with pages and get recommendations based on the clusters in which the pages exist.

To evaluate and explore the interconnected nature of Wikipedia articles and simulate how users might engage with this information, several experiments were conducted using graph-based and embedding-based methods. The core of this experimentation focused on understanding user interest similarity through article recommendations and providing intelligent pairing of users for collaborative learning.

A group of synthetic users was generated to simulate real students browsing and exploring Wikipedia. Each user was assigned a list of recommended Wikipedia pages, selected based on a seeded starting page and a graph traversal method that prioritizes relevance and diversity. These recommendations were derived from a larger graph structure

constructed using NetworkX, where nodes represent Wikipedia pages and edges represent hyperlinks between them. The traversal ensured that users would receive sets of articles that varied in size and topical coverage, simulating diverse academic interests.

To assess how similar users interests were, we used a pre-trained Sentence Transformer model (all-MiniLM-L6-v2) from the sentence-transformers library. Each users recommended article list was converted into titles, which were then encoded into high-dimensional vector representations (embeddings). The embeddings capture semantic information about each article, allowing us to compare articles beyond just their titles or keywords.

To compute pairwise similarity between users, we used the average cosine similarity between the sentence embeddings of their recommended articles. This involved calculating all pairwise similarities between the article embeddings of two users, then averaging them to obtain a single similarity score. These scores allowed us to rank how closely aligned each users interests were with others.

To improve performance on Colab and support GPU acceleration, the model was loaded using device='cuda', and inference operations were optimized using CuPy where possible. However, the memory demands of full pairwise comparisons with large embeddings led to high RAM usage, prompting additional exploration of lightweight similarity measures.

In addition to semantic comparison, we evaluated user similarity using the Jaccard index, which measures the overlap between two sets. For this experiment, the recommended article lists for each user were treated as sets, and Jaccard similarity was computed based on shared page IDs. This provided a faster and more memory-efficient way to approximate user similarity, especially in cases where exact matches were more relevant than semantic ones.

## 5. RESULTS

For every user, we can generate recommended articles based on the visited pages that the user has interacted with. These are some example generated from an example users visited pages and its recommended wikis for the user.

### Summary of Recommended Articles to look into

#### --- The Curse of the Yellow Snake ---

The Curse of the Yellow Snake (German: Der Fluch der gelben Schlange) is a 1963 West German crime thriller film directed by Franz Josef Gottlieb and starring Joachim Fuchsberger, Brigitte Grothum and Pinkas Braun. It is based on the 1926 novel The Yellow Snake by Edgar Wallace. It was made as part of a series of films based on Wallace's work, made either by CCC Film or the rival Rialto. It was shot at the Spandau Studios in West Berlin and on location in London. The film's sets were designed by the art directors Hans Jrgen Kiebach and Ernst Schomer.

#### --- The Curse of the Hidden Vault ---

The Curse of the Hidden Vault (German: Die Gruft mit dem Rtselschlo) is a 1964 black and white West German crime film directed by Franz Josef Gottlieb and starring Harald Leipnitz, Eddi Arent, Siegfried Schrenberg and Klaus Kinski. It is based on the 1908 novel Angel Esquire by Edgar Wallace, previously made into a British silent film. It was shot at the Spandau Studios and Tempelhof Studios in Berlin and on location in London. The film's sets were designed by the art directors Wilhelm Vorwerk and Walter Kutz.

#### --- Edgar Wallace ---

Richard Horatio Edgar Wallace (1 April 1875 10 February 1932) was a British writer of crime and adventure fiction. Born into poverty as an illegitimate London child, Wallace left school at the age of 12. He joined the army at age 21 and was a war correspondent during the Second Boer War for Reuters and the Daily Mail. Struggling with debt, he left South Africa, returned to London and began writing thrillers to raise income, publishing books including The Four Just Men (1905). Drawing on his time as a reporter in the Congo, covering the Belgian atrocities, Wallace serialised short stories in magazines such as The Windsor Magazine and later published collections such as Sanders of the River (1911). He signed with Hodder and Stoughton in 1921 and became an internationally recognised author.

```

110 --- The Strange Countess ---
111 The Strange Countess (German: Die seltsame Grfin) is a 1961 West German crime film directed by Josef von
112 Bky and starring Lil Dagover, Joachim Fuchsberger and Marianne Hoppe. It is based on Edgar Wallace's
113 1925 novel of the same title, and is part of a long-running series of Wallace adaptations produced by
114 Rialto Film. It was shot at the Tempelhof Studios in Berlin. Location shooting took place at the
115 Schloss Ahrensburg. The film's sets were designed by the art director Helmut Nentwig.
116 -----

```

118 In the end we want to be able to create heatmaps for users to see which students are studying the same topics  
119 usding the Jaccard similarity. While currently through testing the Jaccard similarities computed don't show any high  
120 similarity to any of the synthetic users, we can still visualize the concept we are trying to grasp by using a heat map.

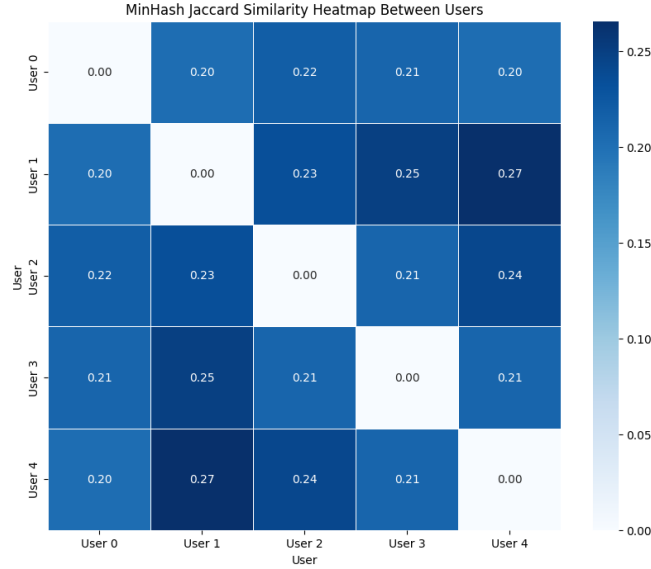


Figure 5. Jaccard Similarity Heatmap

## 6. FURTHER WORK

122 As shown in Figure 5, users 1 and 4 received identical sets of recommended articles from the knowledge graph-based  
123 recommendation system. This suggests that they may be good candidates to study together. However, while the figure  
124 visually indicates this connection, the Jaccard similarity score provides a more rigorous measure of overlap between  
125 user recommendations. In our analysis, we observed that meaningful similarity tends to emerge when the Jaccard  
126 index exceeds a threshold of approximately 0.65.

127 A key challenge here lies in the randomness of the recommendation generation process. Since the article suggestions  
128 are drawn based on sampling indices from a Gaussian distribution, repeated simulations are necessary to achieve  
129 convergence in user similarity scores. With enough iterations, users who share genuine topical interest are more likely  
130 to receive overlapping recommendations. However, this requires more computational resources, which currently limits  
131 our ability to simulate large-scale user interactions. Overcoming this limitation would allow us to simulate a greater  
132 number of users and visualize a more complex and nuanced network of user interactions.

133 Additionally, one limitation of relying solely on Jaccard similarity is that it compares the presence or absence of  
134 items in each users recommended set, without considering the actual meaning or content of those articles. To address  
135 this, we also experimented with semantic similarity using Sentence Transformers. This approach involved computing  
136 cosine similarity between sentence embeddings of recommended article titles, allowing us to assess the conceptual  
137 overlap between user interests even when the exact articles did not match. While this method is more computationally  
138 intensive, it offers a deeper and more flexible understanding of how users relate in terms of what they are learning.