Wiki Insights: Extracting Personalized Learning Path From Wikipedia Pages

Nikki Saggi, Linnea May, Ziwei Gu, Xinqi Lyu, Jim Z. Li

Introduction

Wikipedia, one of the most visited web-based encyclopedia sites, has the largest collections of knowledge contents curated by collaborators worldwide. The user experience with finding the right contents on Wikipedia, however, is challenging. Users likely need to read through multiple pages to understand a concept, without any idea about the most intuitive reading order.

Our team is aspired to create a web application that recommends users the best possible reading order of Wikipedia pages for a concept they want to understand. This application takes in the hyperlink network structure of Wikipedia pages, topics and contents similarities, and random walks to generate a suggested reading order.

Hypotheses

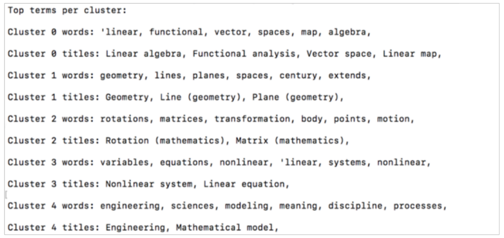
We formulated two hypotheses based on previous research:

H0: There exists a strictly hierarchical structure of concepts for any given subject.

H1: There exists a general hierarchy of abstraction from which a loose path of learning can be built.

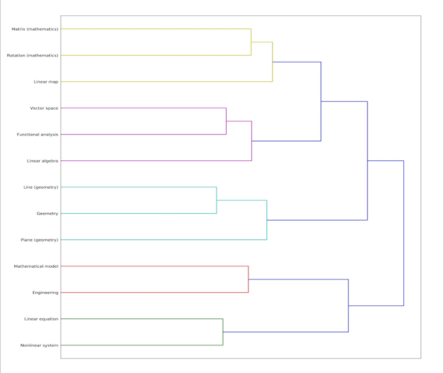
Approach

Several clustering techniques were taken for extracting the commonalities between Wikipedia pages and intrinsic topical structure.



```
Top terms per cluster:

Cluster 0 words: 'linear, functional, vector, spaces, map, algebra,

Cluster 0 titles: Linear algebra, Functional analysis, Vector space, Linear map,

Cluster 1 words: geometry, lines, planes, spaces, century, extends,

Cluster 1 titles: Geometry, Line (geometry), Plane (geometry),

Cluster 2 words: rotations, matrices, transformation, body, points, motion,

Cluster 2 titles: Rotation (mathematics), Matrix (mathematics),

Cluster 3 words: variables, equations, nonlinear, 'linear, systems, nonlinear,

Cluster 3 titles: Nonlinear system, Linear equation,

Cluster 4 words: engineering, sciences, modeling, meaning, discipline, processes,

Cluster 4 titles: Engineering, Mathematical model,
```

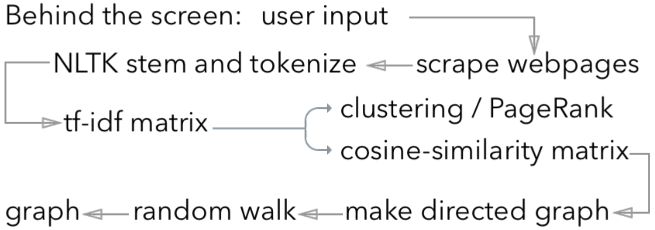
K-Means Clusteirng



Hierarchical Clusteirng

However, topic modeling does not reflect the hyperlink structure of the interconnectedness of Wikipedia pages and concepts. We took to constructing a directed graph with edges weighted by cosine similarities.

Random walks were performed to find a converging path.



Behind the screen: user input

scrape webpages

NLTK stem and tokenize

tf-idf matrix

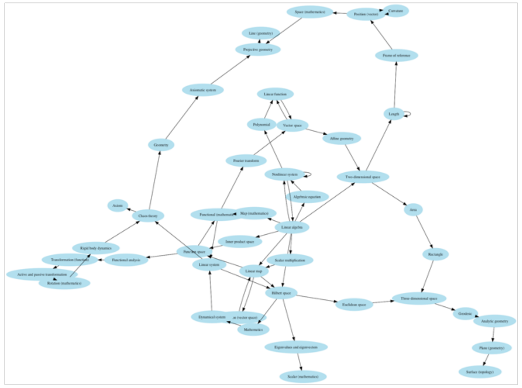
clustering / PageRank

cosine-similarity matrix

make directed graph

random walk

graph



Future Goals

1. Experiment with different weighting factors, such as co-editor index between two pages.

2. Tune cosine similarity value threshold to improve results.

3. Expand query to non-structural concepts.