Wiki Insights: Extracting Personalized Learning Path From Wikipedia Pages

Nikhil Saggi, Ziwei Gu, Xinqi Lyu, Eric Sun, Debasmita Bhattacharya, Ellen Chen

Cornell University

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 necessary preliminary knowledge or related topics 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 inspired 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

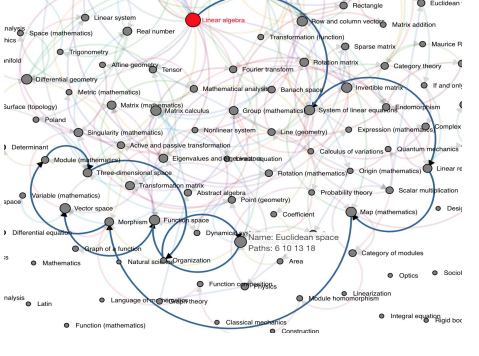
We formulated two hypotheses based on previous research:

- **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

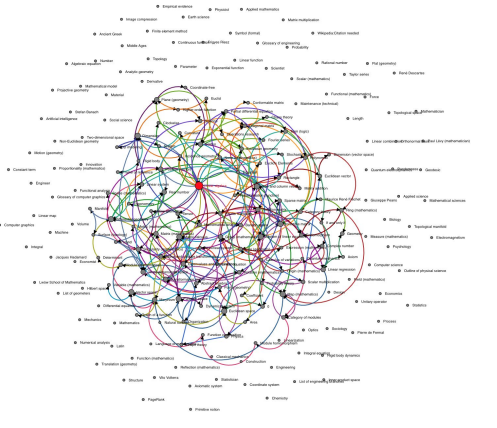
1. **Topic Modeling:**
   - Building a word dictionary and term-document matrix from corpus
   - Leveraging the internal structure of Wikipedia: Infobox, Category, etc.
   - Clustering documents based on the Latent Dirichlet Allocation (LDA) Model
   - Creating similarity matrix from a combined Latent Semantic Indexing (LSI) and TF-IDF model

2. **Building a Wikipedia Graph:**
   - Articles as vertices, links as edges
   - Links factored out according to generality; target (t) more general than source (s) if #inlink(t) / #inlink(s) ≥ k
   - Defining an ideal graph: frequency of self loops, degree of centrality, average depth of branches.



3. **Random Walk:**
   - At each node, a weighted random step was made from edges leaving the node.
   - Path taken based on the weight of the edges.
   - Walk proceeds towards more specific or more general topics, contingent on whether nodes with more or less in-links are prioritized in the step calculation.

Final Visualization



References

E. Yeh, D. Ramage, C.D. Manning, WikiWalk: Random walks on Wikipedia for Semantic Relatedness (2016)
E. Minkov, W. Cohen, Learning Graph Walk Based Similarity Measures for Parsed Text (2010)