WikipediaとLLMを使った知識マップの構築

SciPyData2025（1/25, 2025）

辻真吾（www.tsjshg.info）

自己紹介

- Pythonを使ったデータサイエンスが得意
  - バイオインフォマティクス
  - エネルギーシステム
  - 学習支援
- 最近はRustを書きたい
- RISC-Vに興味がある
- 所属など
  - 東京大学先端科学技術研究センター・先端データサイエンス分野
  - アークエルテクノロジーズ株式会社 CAIO
  - 株式会社RATH 技術アドバイザー
- Python、データサイエンス、アルゴリズムに関する著書多数
- https://www.tsjshg.info

もくじ

- 研究の背景と目的
- アプリケーションの紹介
- 周辺技術の解説
- 今後の課題と展望

背景

- AI・データサイエンスに関する知識を中心に学ばなければならない項目が急増している
- 初学者にとって、自分の立ち位置を把握し、身につけたい知識への学習パスを把握することは困難

目的

今持っている知識とゴールを入力すると、学習パスを提示してくれるシステムの開発

アプリケーションを作ったのでその実行例をお見せします

Demo1

Deploy

Configuration

already known items

machine learning
neural network

your goals

ChatGPT

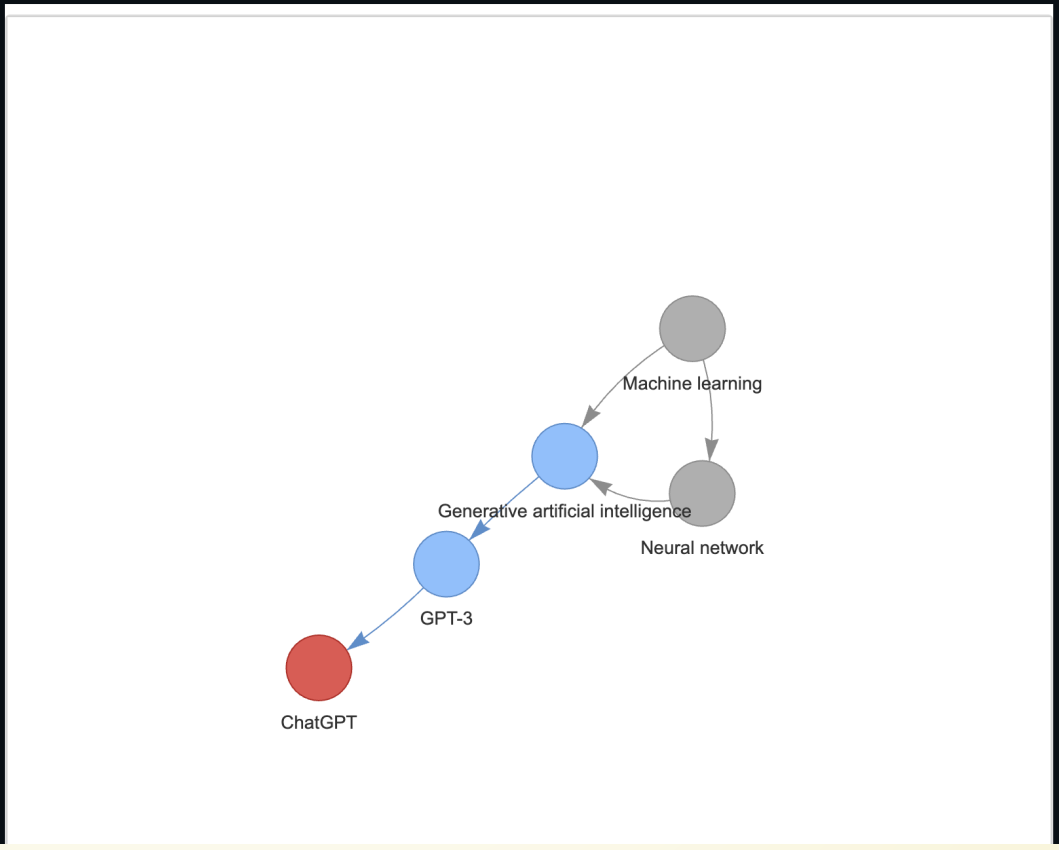
cos similarity threshold

0.95

0.70 1.00

Submit

Your optimized learning path is here.



Demo2

Deploy

Configuration

already known items

sigmoid function

your goals

deep reinforcement learning

cos similarity threshold

0.95

0.70 1.00

Submit

Your optimized learning path is here.

Demo3

Deploy

Configuration

already known items

decision tree (tree data structure)

your goals

random forests

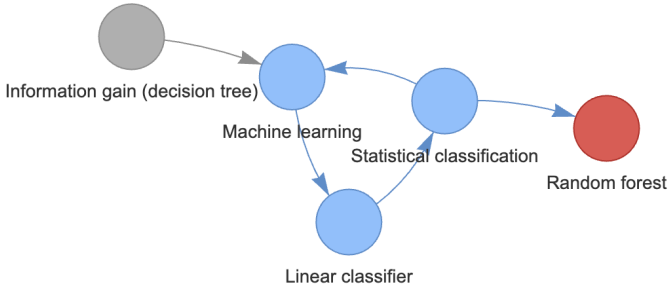
cos similarity threshold

0.95

0.70 1.00

Submit

Your optimized learning path is here.



知識マップ

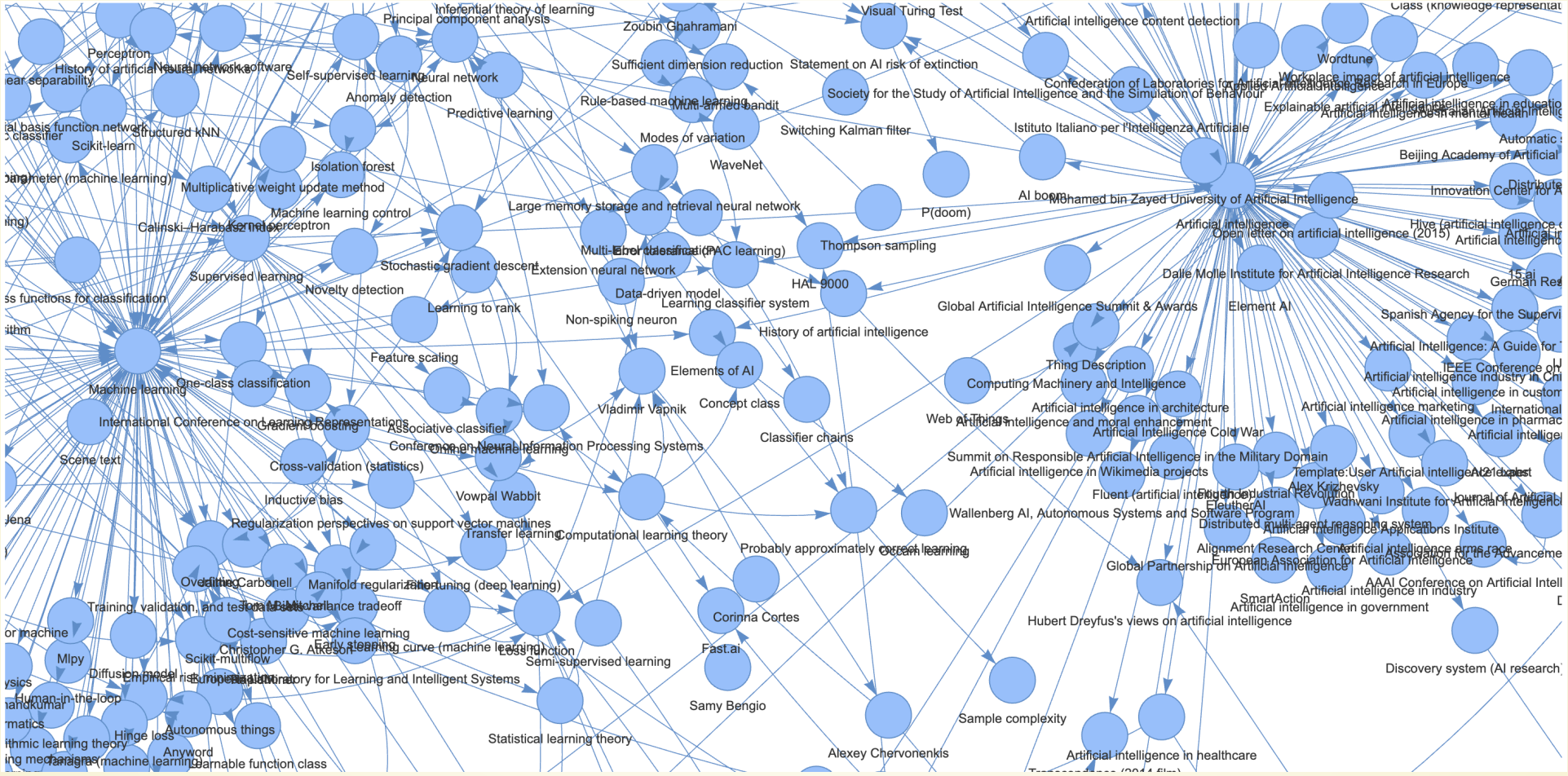
- 学習項目の関係を保存する有向グラフ
- 英語版Wikipediaのデータを利用
  - グラフのノードはWikipediaの1ページ
- Artificial Intelligenceカテゴリーの関連ページを取得
- ページ間のリンクの依存関係をChatGPT-4oへ問い合わせる

ChatGPT-4o

- XがYの基礎項目なら1、XがYの発展項目なら2、無関係な0と答えてもらう
- Wikipediaのページのタイトルと冒頭説明をセットで入力
- XとYの順序を入れ替えたとき答えの整合性がとれるものだけに限定
- Xの発展がYのときX→Yと表記する

知識マップの全体像

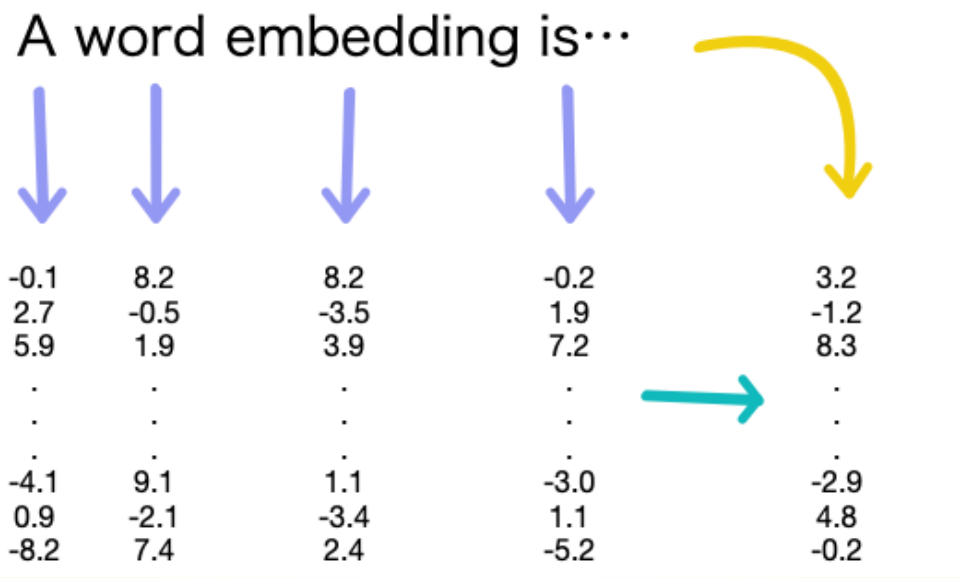
- 最大の連結成分
- 1,568ノード、2,358エッジ



問題点と課題

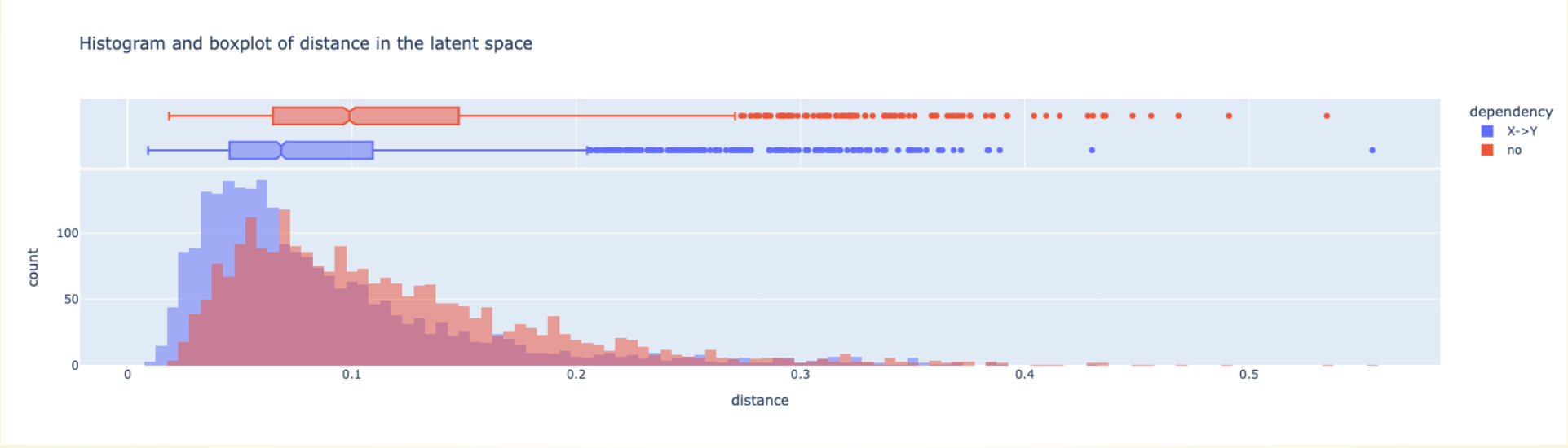
- X,Yの順に渡すとXを基礎項目と答える傾向が強い（4倍以上）
- 人手による精度の見積もり
- プロンプトエンジニアリングを頑張るか新バージョンに期待するか
  - 計算資源があれば独自モデルの構築もあり得るか？

単語の埋め込み（word embeddings）



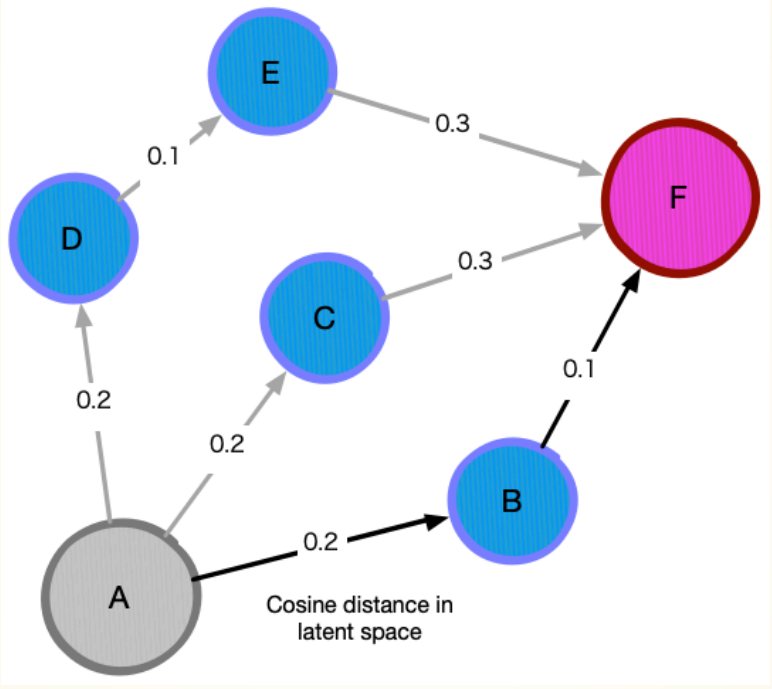
- 潜在空間におけるドキュメント検索では、全体をまとめたベクトルが使われることが多い（図右下）
- Wikipediaの冒頭をトークンごとのベクトル化して検索

X→Yの関係と潜在空間での距離



- 知識マップにエッジがある2つの項目は潜在空間で近くに存在する傾向がある
- 広大な潜在空間からLLMを使ってなんらかの関係性を引き出せている

学習パス

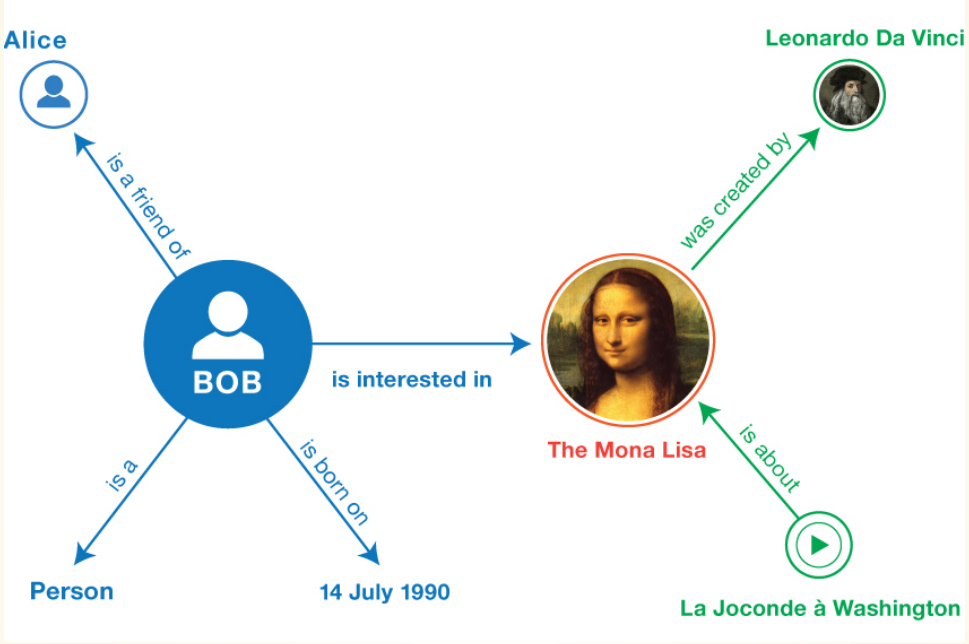


- 潜在空間内での距離を考慮した最短経路
  - 有向道が見つからないときは向きを無視した道
- これはいろいろ改良の余地がありそう

関連技術の紹介

- 知識マップの生成
  - RDF（resource description framework）と知識グラフ
- 情報検索
  - Oracle社の新サービス
- 学習パスの探索
  - GNN（graph neural network）を使った例

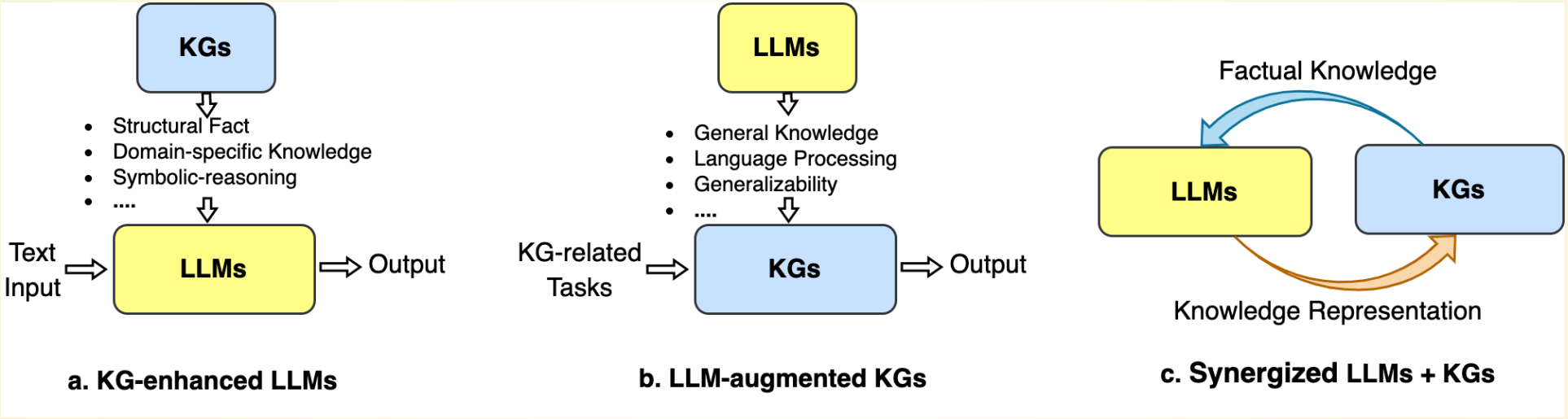
知識グラフ



- 主語、述語、目的語の3つ（トリプル）で情報を表現
- トリプルをつなげてグラフを構築

RDF1.1入門　https://www.asahi-net.or.jp/~ax2s-kmtn/internet/rdf/NOTE-rdf11-primer-20140225.html

知識グラフとLLM



- これまでの知識グラフ研究をLLMで拡張、発展させる研究が出てきている

検索とRAG (retrieval augmented generation)

Artificial Intelligence

Oracle Announces General Availability of AI Vector Search in Oracle Database 23ai

May 2, 2024 | 5 minute read

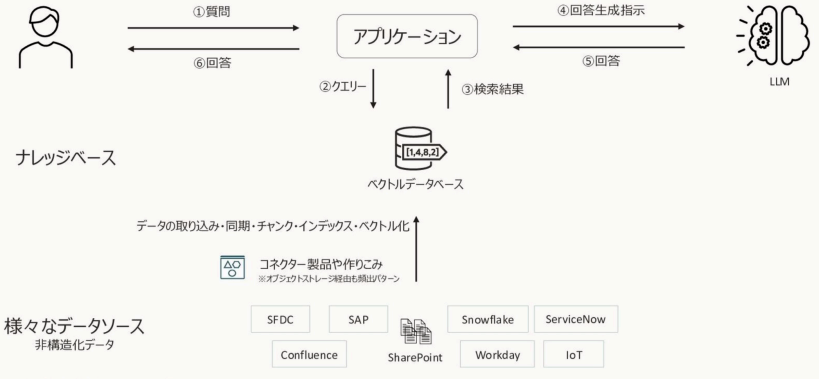

Doug Hood
Oracle AI Vector Search Product Manager


Ranjan Priyadarshi
Senior Director, Product Management(Mission-Critical Data and AI Engines)

Oracle AI Vector Search is a novel capability that allows users to search data based on the semantics or meaning of data, in addition to by the values of data, such as attribute values or keywords, as databases have traditionally supported.

A vector, or vector embedding, is a popular data structure used in AI applications. A vector is a list of numbers, generated by deep learning models from diverse data types (e.g. images, documents, videos, etc.), that encodes the semantics of the data.

RAG のアーキテクチャ



左Oracle Database Insiderから引用 右Oracle AI Vector Search 技術概要から引用

Graph Neural Networkを使った学習パスの同定

Personalized Learning Path Generation Algorithm Based on Graph Neural Networks

Conference paper | First Online: 13 October 2024

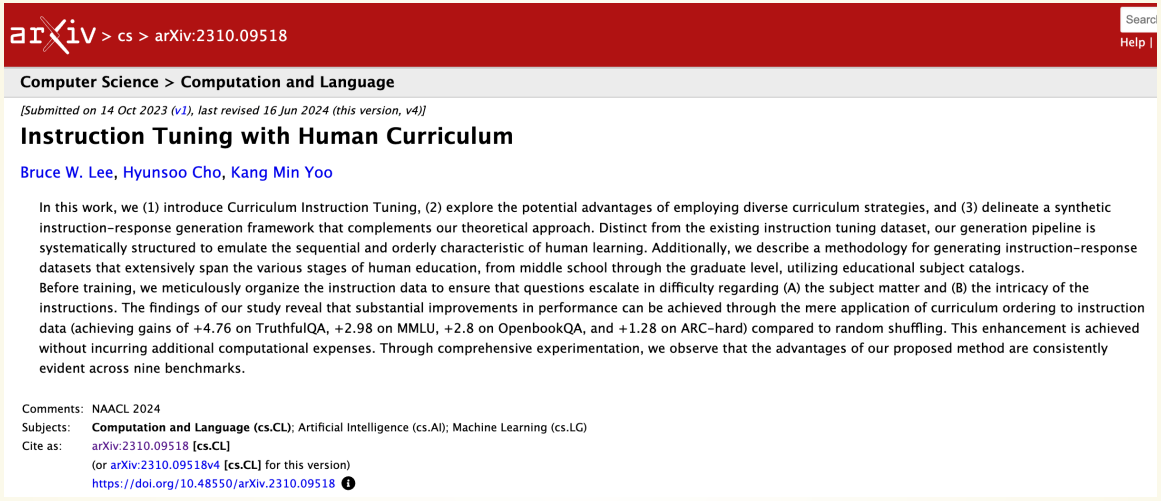
pp 281–291 | Cite this conference paper

Abstract

With the rapid development of information technology and educational resources, traditional learning path recommendation systems are often based on students' historical data or resource categorization, making it difficult to fully meet the personalized needs and ability differences of students. This study proposes a personalized learning path generation algorithm based on Graph Neural Networks (GNNs), which can recommend the most suitable learning paths for students based on their abilities, interests, and learning needs, thereby improving learning efficiency and optimizing the allocation of educational resources. This paper first constructs the model by introducing the concept and application of GNNs, then conducts experimental comparisons with content recommendation algorithms and collaborative filtering algorithms, and finally validates the effectiveness of the proposed algorithm on a specific dataset. The experiments are conducted on a dataset of 50 students from an internet education platform. The model training utilizes students' behavioral data, interests, hobbies, and study time for training and fine-tuning. The experimental results show that the algorithm proposed in this study outperforms traditional content recommendation and collaborative filtering algorithms in terms of recommendation accuracy, recall rate, and F1 score. This research provides theoretical and practical guidance for the design of personalized learning paths in the direction of online education.

Personalized Learning Path Generation Algorithm Based on Graph Neural Networksから引用

LLMを使った学習パス構築の可能性



arXiv > cs > arXiv:2310.09518

Search
Help |

Computer Science > Computation and Language

[Submitted on 14 Oct 2023 (v1), last revised 16 Jun 2024 (this version, v4)]

Instruction Tuning with Human Curriculum

Bruce W. Lee, Hyunsoo Cho, Kang Min Yoo

In this work, we (1) introduce Curriculum Instruction Tuning, (2) explore the potential advantages of employing diverse curriculum strategies, and (3) delineate a synthetic instruction-response generation framework that complements our theoretical approach. Distinct from the existing instruction tuning dataset, our generation pipeline is systematically structured to emulate the sequential and orderly characteristic of human learning. Additionally, we describe a methodology for generating instruction-response datasets that extensively span the various stages of human education, from middle school through the graduate level, utilizing educational subject catalogs.
Before training, we meticulously organize the instruction data to ensure that questions escalate in difficulty regarding (A) the subject matter and (B) the intricacy of the instructions. The findings of our study reveal that substantial improvements in performance can be achieved through the mere application of curriculum ordering to instruction data (achieving gains of +4.76 on TruthfulQA, +2.98 on MMLU, +2.8 on OpenbookQA, and +1.28 on ARC-hard) compared to random shuffling. This enhancement is achieved without incurring additional computational expenses. Through comprehensive experimentation, we observe that the advantages of our proposed method are consistently evident across nine benchmarks.

Comments: NAACL 2024
Subjects: **Computation and Language (cs.CL)**; Artificial Intelligence (cs.AI); Machine Learning (cs.LG)
Cite as: arXiv:2310.09518 **[cs.CL]**
(or arXiv:2310.09518v4 **[cs.CL]** for this version)
https://doi.org/10.48550/arXiv.2310.09518

- 人が学ぶのと同じカリキュラムの順でLLMを鍛えると性能が向上するという報告
- もしこれが事実なら、さまざまなlearning pathの評価をLLMエージェントにさせることで学習パスを最適化できる可能性がある

Instruction Tuning with Human Curriculumから引用

まとめ

- Wikipediaのデータを例に、学習パスを提示するソフトウエアを開発し、その周辺技術を紹介
- 知識マップの精度向上は大きな課題
  - 知識や概念の階層性など

謝辞

この研究はBeyondAI研究推進機構を通じてソフトバンクから資金提供を受けています

研究協力：（株）CMSコミュニケーションズ、（株）ゆめみ