Your Thesis Title

Submitted September 2023, in partial fulfillment of
the conditions for the award of the degree **MSc Computer Science.**

Yu Hao
20498393

Supervised by Kai Xu

School of Computer Science
University of Nottingham

I hereby declare that this dissertation is all my own work, except as indicated in the text:

Signature Yu Hao

Date 08 / 09 / 2023

I hereby declare that I have all necessary rights and consents to publicly distribute this dissertation via the University of Nottingham's e-dissertation archive.

Abstract

The "Projection of Provenance Vector Sequence" project represents an innovative approach to enhance the understanding of data lineage and history within complex information systems. In this endeavor, a sequence of vectors, encapsulating the provenance of data, is meticulously analyzed and visualized. This endeavor primarily aims to decipher the intricate relationships and transformations that underlie data evolution. Leveraging advanced techniques in data visualization and analysis, the research explores the creation of clear, informative visualizations that enable users to grasp the historical context of data. Additionally, novel strategies for vectorization and quantization of data provenance are introduced, offering valuable insights into the evolution of data states.

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor, Kai Xu, for providing me with a lot of patience and practical advice during the project process. Secondly, I would like to thank my three teammates, Xinhao Yang, Meirong Li, and Yuchen Zhou, who have provided me with assistance in the LangChain model, visualization design, and related knowledge respectively. Also, I am very grateful to the developers of the software and tools used in this project.

Contents

List of Tables

List of Figures

Chapter 1

Introduction and Background

1.1 Background

In an era where data has emerged as a driving force behind decision-making, innovation, and societal progress, understanding the provenance of information has gained substantial significance. Provenance encapsulates the origin, history, transformations, and relationships of data, providing critical insights into how information evolves over time. As data ecosystems become more intricate, the need to navigate and comprehend these complex journeys has led to the emergence of innovative techniques such as the "Projection of Provenance Vector Sequence."

The concept of provenance traces its roots to various domains, including data management, science, and art authentication. In data management, provenance enables tracking the lineage of data transformations, enhancing data quality and traceability. In scientific research, provenance is essential for ensuring reproducibility and transparency in experiments and analyses. Similarly, the art world employs provenance to authenticate artworks and establish their historical lineage. However, as information systems and processes have evolved, traditional methods of representing and visualizing provenance have encountered limitations, particularly when dealing with dynamic, multidimensional, and intricate data sequences.

The "Projection of Provenance Vector Sequence" emerges as a response to the challenges of visualizing complex information pathways. This technique capitalizes on advancements in vector-based representations and data visualization to create a dynamic visual representation of provenance vectors. By projecting these vectors into a visual space, the technique offers a unique lens through which to explore the evolution of data, information flow, and the interplay of concepts and entities over time.

Drawing inspiration from fields like information visualization, network analysis, and computational linguistics, the ”Projection of Provenance Vector Sequence” aims to bridge the gap between intricate data sequences and human cognition. This approach acknowledges that while data may inherently reside in a high-dimensional space, human perception is more attuned to visual patterns and relationships. By projecting these abstract vectors into a visual domain, the technique empowers users to discern patterns, trends, and anomalies that could have significant implications in domains ranging from sensemaking and decision support to scientific discovery and collaborative research.

The rise of this technique also aligns with the growth of the data-driven economy, where the ability to interpret and navigate complex data landscapes directly influences competitive advantage, innovation, and informed decision-making. As the world generates an ever-increasing volume of data, the ”Projection of Provenance Vector Sequence” offers a promising avenue to not only enhance our understanding of data evolution but also to empower individuals to harness the power of data-driven insights more effectively.

In the rapidly evolving landscape of data-driven decision-making, understanding the intricate journeys of information, knowledge, and data flow has become paramount. The ”Projection of Provenance Vector Sequence” emerges as a novel approach to tackle the complexities of information provenance, offering a visual gateway into the otherwise cryptic pathways of data evolution. By leveraging advanced visualization techniques and vector-based representations, this innovative methodology endeavors to illuminate the dynamics of information dissemination, enabling users to decipher patterns, interconnections, and transformations within vast data sequences.

1.2 Motivation

In a world inundated with data, the need to trace the origin, transformation, and impact of information has never been more crucial. However, deciphering these multifaceted data narratives poses a substantial challenge, often hindered by the sheer complexity of information ecosystems. The "Projection of Provenance Vector Sequence" is motivated by the quest to bridge this gap by translating intricate data trails into coherent visual narratives. This not only empowers data professionals, researchers, and decision-makers to gain insights with unprecedented clarity but also opens doors for collaborative sensemaking, informed decision-making, and more profound comprehension of complex information processes.

1.3 Aims and Objectives

The primary objectives of this approach are twofold: to develop a robust framework for projecting provenance vector sequences onto visual representations and to empower users with a tool that facilitates the exploration of complex data trajectories. Through this, the technique aims to:

- **Uncover Hidden Insights**: One of the core objectives of this technique is to uncover latent patterns and trends within complex data sequences. By projecting provenance vectors onto a visual representation, the technique aims to highlight subtle relationships, periodic behaviors, and emerging trends that might otherwise remain concealed within the intricate data landscape. This enables analysts, researchers, and decision-makers to extract insights that can inform strategies, optimizations, and proactive interventions.

- **Facilitate Informed Decision-Making**: The technique aspires to enhance the decision-making process by providing decision-makers with a visual narrative of how information evolves and impacts outcomes. By mapping the provenance vectors onto intuitive visual displays, the technique empowers stakeholders to comprehend the consequences of different choices, anticipate potential outcomes, and make informed decisions that align with their objectives.

- **Enhance Exploratory Data Analysis**: Exploratory Data Analysis (EDA) is a fundamental step in data analysis workflows. The technique aims to augment EDA by providing researchers with a versatile tool to interactively navigate through data sequences, and adjust visualization parameters and methods. This empowers researchers to identify outliers, anomalies, and unexpected behaviors that might hold crucial implications for their investigations and further research.

- **Enhance Data Communication**: The visualization of provenance vectors bridges the gap between data and meaningful communication. The technique seeks to enhance data communication by translating abstract data sequences into visual narratives that resonate with stakeholders. Whether presenting to technical or non-technical audiences, the technique empowers communicators to convey the story of data evolution effectively.

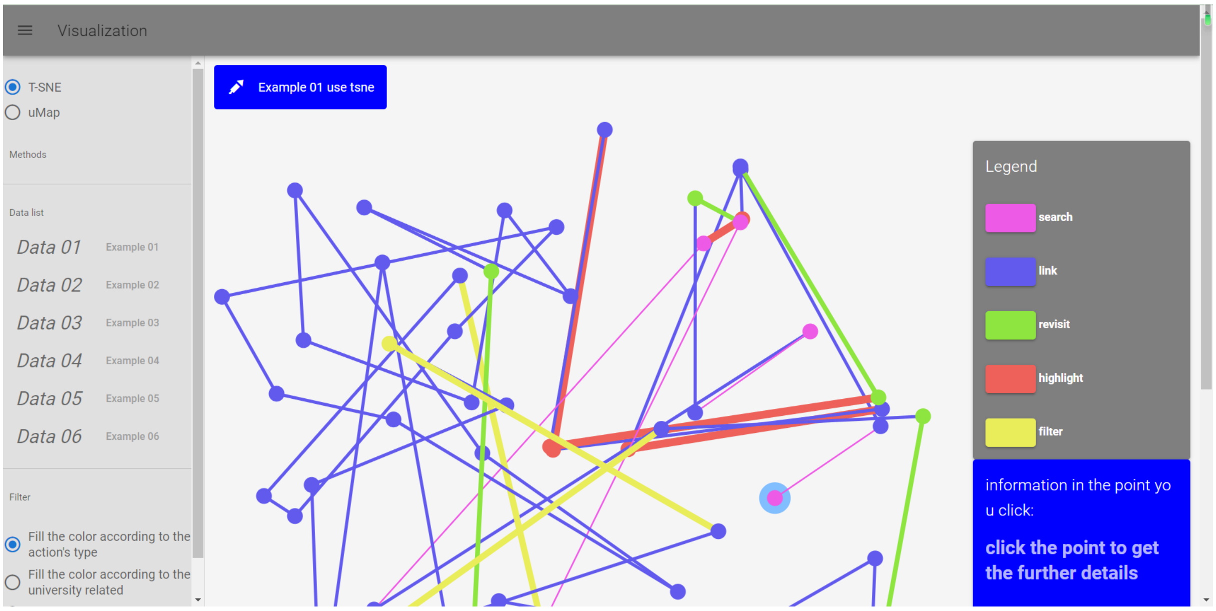
- **Inform Strategy Formulation**: Strategy formulation in domains such as business, policy, and research relies on comprehensive insights. The technique contributes to strategy formulation by providing stakeholders with a holistic view of data trajectories. This aids in identifying leverage points, optimizing resource allocation, and designing strategies that account for the complex interplay of data and information flow.

1.4 Description of the work

The journey of the ”Projection of Provenance Vector Sequence” technique involves a multi-faceted approach. It begins with the formulation of effective methods to represent complex provenance vectors in a visual format. Advanced visualization tools, such as using Vue to design the user interface including network graphs and interactive displays, are harnessed to translate abstract vectors into meaningful visual constructs. The project further delves into designing user-friendly interfaces that allow users to interact with the visualization, zooming into specific tasks, exploring relevant information, and discerning connections between data nodes. Additionally, the technique will integrate feedback loops to refine the visual representations based on user experience and needs in future research, ensuring that the resulting visualizations are both informative and intuitive.

In conclusion, the ”Projection of Provenance Vector Sequence” presents an exciting avenue for unraveling the intricacies of data evolution through the marriage of advanced data representations and visual storytelling. By striving to make the invisible pathways of information flow visible, this innovative approach promises to reshape how we perceive, analyze, and make sense of the complex tapestry of data in an increasingly interconnected world.

For my project, I design a web project of visualization of two-dimension projection of provenance vector sequences, which represents the user's sensemaking process. The plot shows that what the process is, how the process works and what the final result is. Meanwhile, the researchers can use this web project as a tool to analyze the user's behaviour and probably predict the user's future action. The project is shown below:



● T-SNE
○ uMap

● Fill the color according to the action's type
○ Fill the color according to the university related

information in the point yo u click:
click the point to get the further details

Figure 1.1: web page of the project

Chapter 2

Related Work

2.1 Literature Review

2.1.1 Visualization and sensemaking

The report in 2017 pointed that While progress has been made, the tactful combination of machine learning and data visualization is still under-explored. Further, it presents opportunities and challenges to enhance the synergy between machine learning and visual analytics for impactful future research directions.[5] Kai Xu, Alvitta Ottley, Conny Walchshofer, Marc Streit, Remco Chang and John Wenskovitch emphasized the importance of SURVEY in their research, they aimed to produce a comprehensive survey of work in the data visualization and visual analytics field that focus on the analysis of user interaction and provenance data. Therefore, they creatively structured their survey around three primary questions: (1) WHY analyze provenance data, (2) WHAT provenance data to encode and how to encode it, and (3) HOW to analyze provenance data. [25] Similarly, as for reverse engineering, the report, composed by Wayne C. Henry and Gilbert L. Peterson, introduced SensorRE, the first analytic provenance tool designed to support software reverse engineers, which can automatically capture user's sensemaking actions and provides a graph and storyboard view to support further analysis. [6] Another report on search and rescue pointed out that the use of artifacts like information visualization techniques (e.g., interactive maps) supports sensemaking, allowing teammates to share information, synchronize their activities, and maintain awareness. [1] Guided by sensemaking theory, Gonzalo Ramos and Jina Suh and Rachel Ng and Christopher Meek present ForSense, a browser extension for accelerating people's online research experience. The two primary sources of novelty of ForSense are the integration of multiple stages of online research and providing machine assistance to the user by leveraging recent advances in neural-driven machine reading. [17]

2.1.2 Provenance analysis

Provenance analysis is a widely used technique in various fields. Projecting the provenance vector sequence onto a two-dimensional plane can provide insights into the patterns and trends observed in the data. In this literature review, we discuss the existing research on the processing of provenance vector sequences and highlight the key findings and techniques.

- **PCA**: PCA (Principal Component Analysis) is a commonly used technique for dimensionality reduction and visualization of high-dimensional data. It is used in many fields. For example, in archaeology, Maltsev et al. describe a combination of the total-reflection X-ray fluorescence (TXRF) method and chemometric techniques (PCA, k-means cluster analysis, and SVM) to study Neolithic ceramic samples from eastern Siberia (Baikal region). [11]. In Psychophysiology, to reduce computation time, Pokorny, Victor J., Sponheim, Scott R., Rawls, and Eric decreased EEG dimensionality using PCA prior to ICA, which is called reduced-dimensionality ICA (rdICA), and in some ways, PCA-based rdICA is justifiable when used cautiously. [15]

- **t-SNE**: t-SNE (t-Distributed Stochastic Neighbor Embedding) is a popular algorithm for dimensionality reduction and visualization of high-dimensional data. For example, Walter Serna-Serna et al. used termed Semi-Supervised t-SNE to do the Unsupervised dimensionality reduction and properly fixed the widths of Gaussian neighborhoods to reveal the salient local and global data structures in an low-dimentional space. [18]. Similarly, Yuxian Duan et al. introduced t-SNE to reduce the dimensionality of the original data to solve the redundancy problem caused by excessively high dimensionality. And The Affinity propagation algorithm was optimised on this basuis. [4]

- **UMAP**: UMAP (uniform manifold approximation projection) is a powerful dimensionality reduction and manifold learning technique used in data analysis, visualization, and machine learning. UMAP is particularly well-suited for reducing the dimensionality of high-dimensional data while preserving the underlying structure and relationships within the data. It has gained popularity in various fields, including data visualization, bioinformatics, and natural language processing, for its ability to reveal complex patterns and structures in data. The UMAP algorithm is competitive with t-SNE for visualization quality, and arguably preserves more of the global structure with superior run time performance. [12]

2.1.3 Semantic Textual Similarity

In the vast realm of Natural Language Processing (NLP), the concept of Semantic Textual Similarity (STS) has emerged as a pivotal cornerstone. STS represents a subfield dedicated to quantifying and measuring the likeness or closeness between two distinct pieces of text, such as words, phrases, sentences, or entire documents. Unlike traditional text similarity metrics, which often rely on surface-level features, STS delves deep into the semantic fabric of language, deciphering the contextual and conceptual aspects that underpin textual content.

STS is motivated by the realization that the mere structural resemblance of text can be inadequate in capturing the essence of meaning and context. In contrast, STS strives to unravel the intricate tapestry of language, discerning how words and phrases interconnect and convey nuanced semantics. It is, in essence, an endeavor to imbue computational systems with a more profound understanding of human language.

The applications of STS are manifold and extend across diverse domains. It plays an indispensable role in information retrieval, where it aids in retrieving documents that are conceptually relevant to a user's query. In machine translation, STS assists in determining

the most appropriate translations by assessing the semantic congruity between source and target language texts. Furthermore, STS finds application in sentiment analysis, question answering, summarization, and beyond.
Shaheer et al. shows that higher parameters in a model do not lend to better performance when there is a massive amount of context missing [19]. The word vector techniques based on BERT (Bidirectional Encoder Representations from Transformers) can generate effective word representations from massive text data, and achieves the best classification performance. [8] Usually, BERT needs a massive amount of context, while in this project, the texts to be vectorized are not that huge. Therefore, we need a smaller and faster model instead.

2.2 Relevant tools

2.2.1 Vue

Vue, also known as Vue.js, is a generally used development tool in Front-end development. Compared with traditional front-end development, Vue Can better organize and simplify the web implementation. It is defined as progressive JavaScript framework for building user interface (UI), published in 2014. [26] Vue has three main core features:

1. **Model-View-ViewModel (MVVM)**: The relationship is shown below:

   In figure 2.1, there are three modules: **Model**: model layer, Responsible for handling logic and interacting with the server; **View**: view layer: responsible for converting the data model into UI display, which can be simply understood as an HTML page; **ViewModel**: Connection layer: used to connect the model layer and view layer, serving as a communication bridge between Model and View.[20]

2. **Componentization**: Componentization means to abstract various logic of graphics and non graphics into a unified concept, which is called component, to implement development patterns. In Vue, each .vue file can be regarded as a component. With this feature, it can reduce the coupling degree of the entire system. While keeping

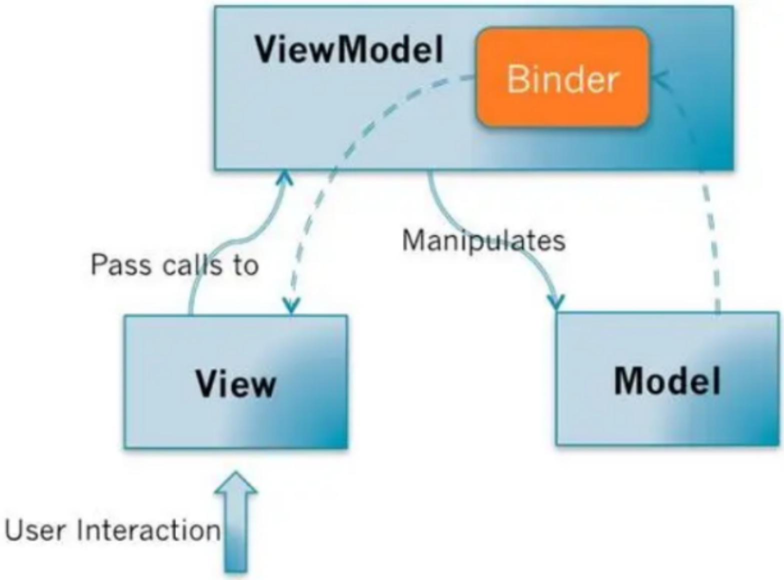


Figure 2.1: Relationship of MVVM

the interface unchanged, we can replace different components to quickly complete the requirements, such as input boxes, which can be replaced with calendar, time, range, and other components for specific implementation. Moreover, Debugging is convenient. As the entire system is composed of components, when problems arise, components can be removed directly using exclusion methods, or problems can be quickly located based on the components that report errors. The reason for fast positioning is that each component has low coupling and a single responsibility, so the logic is simpler than analyzing the entire system. Besides, it will help to improve maintainability, as each component has a single responsibility and is reused in the system, optimizing the code can achieve overall system upgrades.

3. **Directives and DOM**: Directives means special attributes with prefix ”v-” in Vue. DOM (Document Object Model) refers to the structured representation of an HTML document that browsers use to render and manipulate webpages. When the value of an expression in directives changes, its associated effects are applied to the DOM in a responsive manner.

In the report composed by Elar Saks, to compare the framework's performance, a simple single-page application was built in each framework. Applications were then tested in loading speeds. Vue was the fastest performing framework in all tests. In total, Vue was five times faster than Angular and over two times faster than React. Additionally, Vue is Relatively easier to learn.[7]

2.2.2 vuetify

Veutify is an open-source MIT project for web and mobile application front-end development. Vuetify is used in Vue.js framework. For the Vue application, Vuetify is the most extensive UI component library. Vuetify also features a grid system. A grid system is responsible for providing a responsive display for different screen sizes and devices. Besides that, Vuetify has a lot more UI components that can be used in front-end development. [13]

2.2.3 React

React (sometimes called React.js or ReactJS) is an open-source JavaScript library that provides rendering of data into HTML views. React provides programmers with a model where sub-components cannot directly affect outer components (”data flows down”), effectively updating HTML documents when data changes, and providing clean separations among components in modern single-page applications. It encourages the creation of reusable UI components that present data that changes over time. Many people use React as ’V’ in MVC (Model-View-Controller), abstracting the DOM in your DOM, providing a simpler programming model and better performance. Due to these features, React has widely been used in many fields.

In the management field, React.js is regarded as a framework specially designed to deal with the visualization layer of web applications and the author uses it in a business context, namely in human resources, which is the focus of the human portal solution.[2] In the car industry, the author observed how React’s component-based architecture facilitated the development of reusable UI components, resulting in a more modular and maintainable codebase. The use of React’s declarative syntax allowed for efficient rendering of UI elements based on changing data, improving overall performance. [16]

React also has three features:

1. **JSX**: JSX is a JavaScript syntax extension. It is not necessary to use JSX in React development, but it is recommended to use it.
2. **Component**: React is about components. You need to treat everything as a component. This will help you maintain code when dealing with large projects.
3. **Unidirectional data flow and Flux**: React implements one-way data flow, making it easy to infer your application. Flux is a pattern that helps maintain data unidirectionality.

The best part about React is that it uses a declarative style of programming rather than an imperative style. While the former one specifies the compiler what to do, the latter one also has to specify how to do it. Thus, programming with React results in less code. [21]

Chapter 3

User Description

3.1 Target User

The target users for the ”Projection of Provenance Vector Sequence” project are researchers and professionals operating within the domain of data analysis, visualization, and information system management. This user group consists of individuals with diverse backgrounds and expertise, including:

- **Data Scientists**: Researchers who specialize in data analysis, pattern recognition, and information extraction from complex datasets.
- **Information System Analysts**: Professionals responsible for managing, analyzing, and optimizing data flow within information systems.
- **Data Visualization Experts**: Specialists in the design and development of data visualization tools and techniques.
- **Researchers in Data Lineage**: Scholars exploring the intricacies of data lineage, provenance, and historical data transformations.

3.2 User Requirements

To effectively cater to the needs of these researchers, the following user requirements have been identified:

- **Comprehensive Data Understanding**: Researchers require visualization tools that provide a comprehensive understanding of data provenance, enabling them to trace the lineage and evolution of data.

- **Intuitive User Interface**: The tools must feature an intuitive user interface that allows researchers to easily interact with and explore data provenance visualizations. Usability is crucial to ensure efficient workflow.

- **Customization Options**: Researchers often work with diverse datasets and research objectives. Therefore, the tools should offer customization options to adapt to varying research requirements.

- **Accuracy and Reliability**: Researchers demand accuracy and reliability in data representations. The visualization tools must faithfully represent data lineage and transformations to support precise analysis.

- **Task Efficiency**: Efficiency in performing data exploration and analysis tasks is paramount. The tools should facilitate task completion with minimal effort and time.

- **Interactivity**: Interactivity is essential to enable researchers to interact with data visualizations, zoom in on specific details, and extract relevant insights from the provenance data.

- **Scalability**: As researchers often deal with large datasets, scalability is a key requirement. The tools should handle data of varying sizes without compromising performance.

- **Documentation and Support**: Comprehensive documentation and user support resources are essential for researchers to effectively utilize the visualization tools and troubleshoot any issues.

3.3 User Interaction Scenarios

Researchers may interact with the visualization tools in various scenarios, including:

- **Exploring Historical Data Transformations**: Researchers seek to trace the lineage of specific data points, analyze historical changes, and understand how data has evolved over time.

- **Comparing Different Data Lineages**: Researchers compare and contrast data lineages from different sources or contexts to identify patterns, anomalies, or discrepancies.

- **Extracting Insights**: Researchers use the visualization tools to extract actionable insights from data provenance, supporting their research objectives and decision-making.

- **Experimentation and Hypothesis Testing**: Researchers may use the tools for experimentation, hypothesis testing, and validation of research findings related to data lineage and transformations.

3.4 Additional Considerations

Researchers may have varying levels of expertise and familiarity with data visualization and analysis tools. Therefore, the user interface should accommodate both novice and experienced users, providing options for in-depth exploration while maintaining user-friendliness.

Ensuring data privacy and security is paramount, as researchers may be working with sensitive or confidential data. Robust data protection measures should be integrated into the tools to safeguard research data.

As researchers continually seek innovation and advancements in data analysis and visualization, the tools should remain adaptable and capable of accommodating emerging research methodologies and practices.

In summary, the ”Projection of Provenance Vector Sequence” project aims to cater to the specific needs of researchers and professionals in the field of data analysis and data lineage exploration. User requirements encompass usability, customization, accuracy, efficiency, and interactivity, with a focus on supporting diverse research objectives and scenarios.

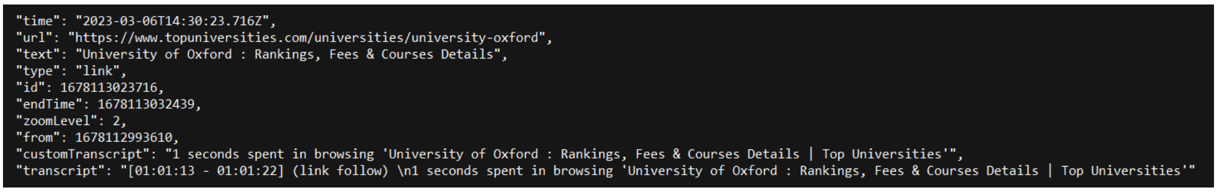
Chapter 4

Methodology

In this project, what to do is to use the data to analyze and obtain the information. Therefore, what methods will be used will be based on the features of the data.

4.1 Data

Figure 4.1 is a group of data used in this project and collected by a Chrome extension called 'SensePath' [14]. With the help of this extension, the data easily can be collected and processed as the form we expected. At the same time, some key information is extracted as well. The data contains several aspects of information, and each group represents a user's action or task. Here is the explanation of each aspect ("(*)" means concerned aspects).



```
"time": "2023-03-06T14:30:23.716Z",
"url": "https://www.topuniversities.com/universities/university-oxford",
"text": "University of Oxford : Rankings, Fees & Courses Details",
"type": "link",
"id": 1678113023716,
"endTime": 1678113032439,
"zoomLevel": 2,
"from": 1678112993610,
"customTranscript": "1 seconds spent in browsing 'University of Oxford : Rankings, Fees & Courses Details | Top Universities'",
"transcript": "[01:01:13 - 01:01:22] (link follow) \n1 seconds spent in browsing 'University of Oxford : Rankings, Fees & Courses Details | Top Universities'"
```

Figure 4.1: Data collected by SensePath

- **(*)"time", "from" and "endTime"**: The start Time and the end time of the action happened, used for calculating the duration of the action.

- **"url"**: The url address of the website.

- **(*)"text"**: The text on the webpage.
- **(*)"type"**: The type of action
- **"id"**: The number used for identify, generated randomly
- **(*)"customTranscript" and "trancript"**: The summary of the action
- **"zoomLevel"**: The scaling of the webpage.

The data is exported as several .json files, while each file is the whole sensemaking process of the user.

4.2 Pre-processing

This part will include three parts: Exploratory Analysis, Vectorization and Dimensionality reduction.

4.2.1 Exploratory Analysis

For our goals, the current form of the data is not sufficient to support our core algorithmic module. Therefore, We need to do some exploratory analysis to make the data meet the requirements of our following processing.

From our analysis of the overall data, we have come to the following conclusions:

- The user conducted a web search for information related to Schools of Computer Science rankings according to the QS World University Rankings and wanted to know some relevant information including fees, courses, etc..
- The user's action can be divided into 5 types: Search, link, revisit, highlight and filter. On the one hand, we can do clustering according to these. While on the other hand, we can infer which action have greater impact on the results. Among all the types, we can easily find "revisit" and "highlight" are mostly the types we are concerned. Due to this, in subsequent processing and analysis, we need to give more weight to both types to make our analysis more precise.

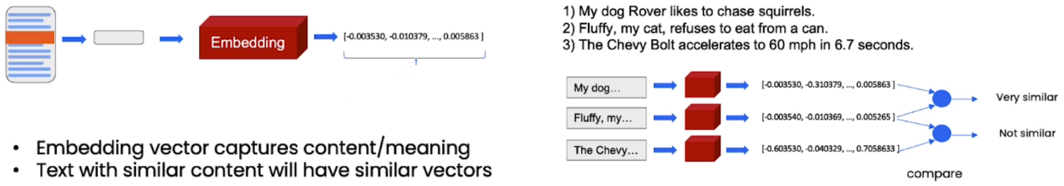
- Other useful indicators are the timestamp and the duration, which can make the user's action organized and traceable.
- The "zoomLevel" and "id" may not use.
- The "text", "customTranscript" and "transcript" is the main part of each action, we will use some model to vectorize them.

4.2.2 Vectorization

Textual content, such as search queries or custom transcripts, can be embedded into dense numerical vectors using techniques like Word2Vec or Doc2Vec. This process captures the semantic meaning of words and phrases.

Here we introduce a model called LangChain. LangChain is an open-source platform for building and deploying natural language processing models [22], including machine translation models.

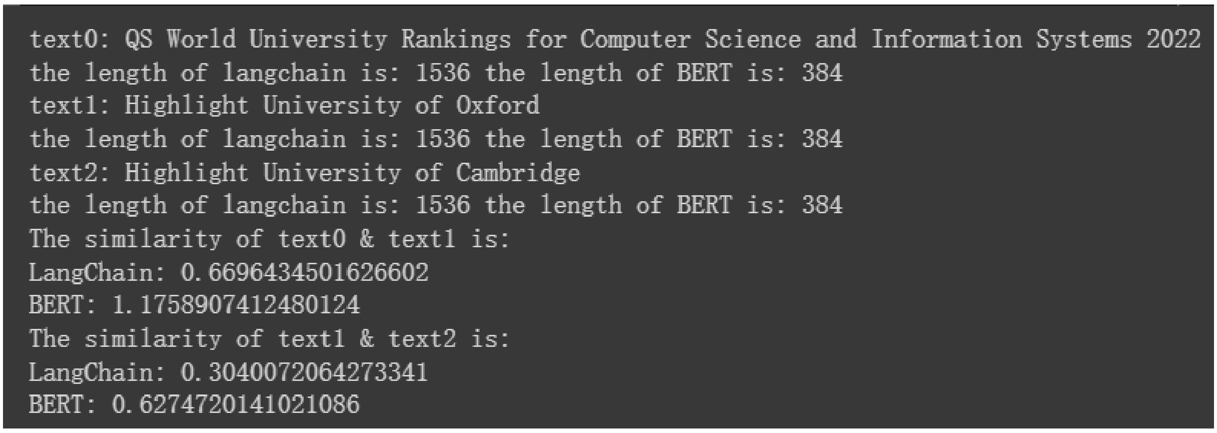
Compared with traditional transformer models, such as Sentence-Transformer, LangChain is a neural machine translation system that uses an encoder-decoder architecture to generate translations from one language to another. Since that, LangChain is typically smaller and faster than BERT as mentioned in 2.1.3. As for vectorization, LangChain uses word embeddings to represent words in a high-dimensional space, while BERT uses a bidirectional transformer architecture that captures both left and right context of a word in a sentence. The embedding layer is one of the key components of LangChain, which is responsible for converting words or phrases into numerical vectors that can be input to a neural network.


Embedding
[-0.003530, -0.010379, ..., 0.005863]
• Embedding vector captures content/meaning
• Text with similar content will have similar vectors
1) My dog Rover likes to chase squirrels.
2) Fluffy, my cat, refuses to eat from a can.
3) The Chevy Bolt accelerates to 60 mph in 6.7 seconds.
My dog...
Fluffy, my...
The Chevy...
[-0.003530, -0.310379, ..., 0.005863]
[-0.003540, -0.010369, ..., 0.005265]
[-0.603530, -0.040329, ..., 0.7058633]
Very similar
Not similar
compare

(a) (b)

Figure 4.2: Structure of embedding layer in LangChain

Figure 4.2(a) shows how the embedding layer works and figure 4.2(b) provides an example of embedding and comparing the similarity of the texts. Obviously, embedding layer performs well in vectorizing short information.



```
text0: QS World University Rankings for Computer Science and Information Systems 2022
the length of langchain is: 1536 the length of BERT is: 384
text1: Highlight University of Oxford
the length of langchain is: 1536 the length of BERT is: 384
text2: Highlight University of Cambridge
the length of langchain is: 1536 the length of BERT is: 384
The similarity of text0 & text1 is:
LangChain: 0.6696434501626602
BERT: 1.1758907412480124
The similarity of text1 & text2 is:
LangChain: 0.3040072064273341
BERT: 0.6274720141021086
```

Figure 4.3: The comparison between LangChain.embeddings and BERT in short information

In figure 4.3, the result shows that LangChain.embeddings will generate more vectors, which means that the description of points is more precise in higher dimension. When comparing the similarity, the result shows that when the texts are similar, LangChain is better, while when the texts are different, BERT is better. Therefore, we can use both of them as alternative or complementary choices.

Besides, some other LLMs (Large Language Model) can do vectorization more precisely and effectively, but require significant GPU memory for inference. [3]

4.2.3 Visualization

The project's primary focus revolves around the visualization techniques and their implementation. To achieve the project's objectives effectively, a combination of visualization methods will be employed. Additionally, the incorporation of graph layout algorithms can significantly enhance the quality of the visualizations. This paper aims to provide a comprehensive overview of the chosen visualization methods and their alignment with the project's goals, with a particular emphasis on the utilization of graph layout algorithms for improved visualization outcomes.

"Provectories" represents a novel approach designed to aid researchers, designers, and

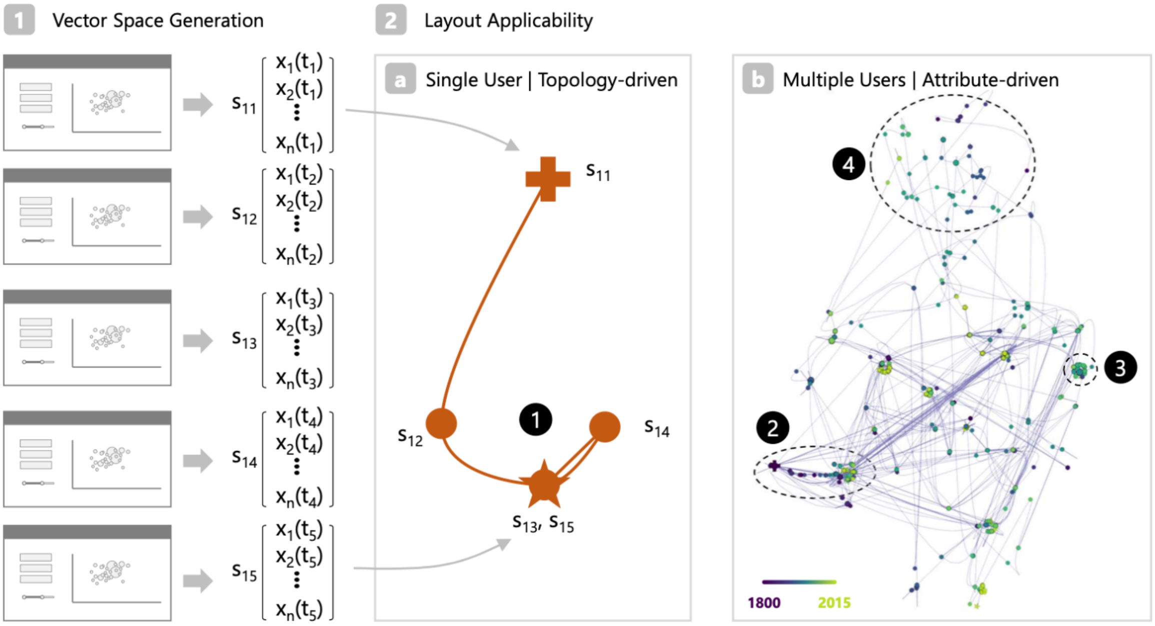

1 Vector Space Generation
2 Layout Applicability
a Single User | Topology-driven
b Multiple Users | Attribute-driven
s_{11}
s_{12}
s_{13}
s_{14}
s_{15}
$x_1(t_1)$ $x_2(t_1)$ $x_n(t_1)$
$x_1(t_2)$ $x_2(t_2)$ $x_n(t_2)$
$x_1(t_3)$ $x_2(t_3)$ $x_n(t_3)$
$x_1(t_4)$ $x_2(t_4)$ $x_n(t_4)$
$x_1(t_5)$ $x_2(t_5)$ $x_n(t_5)$
s_{13}, s_{15}
1
2
3
4
1800
2015

Figure 4.4: Principle of Provectories

developers in the field of visualization. Its primary objective is to enhance the comprehension of user behavioral patterns and analysis strategies. The core concept involves the transformation of recorded application states into feature vectors, subsequently visualized through two distinct layouts: topology-driven and attribute-driven. This paper presents an in-depth exploration of Provectories, including its underlying principles and the introduction of three distinct strategies for the vectorization and quantization of application states. [24] In this project, we are concerned about the 2-dimension projection of those vectors, just like the ④ in figure 4.4.

4.2.4 Dimensionality reduction

In the previous chapter, several methods of dimensionality reduction have been introduced. However, PCA is used for dimensionality reduction by finding orthogonal linear combinations of the original features (principal components) that capture the most variance in the data. While t-SNE and UMAP are both nonlinear algorithms for dimensionality reduction. [23] In this project, our data is real-world data, which means the relationships between data points are not well explained by linear combinations of features and the datasets are less likely to be Gaussian distributed. Due to these reasons, we prefer to use t-SNE and/ or UMAP to reduce dimensionality rather than PCA. And some research has proven that t-SNE and UMAP perform better than PCA in the visualization of the data and preserving the local and global structured data. [12] [10] Both algorithms minimize their loss functions by using gradient descent. Gradient descent begins with some initial configuration of points, and with each iteration, the points are moved to decrease the loss function.[9]

t-SNE

t-SNE constitutes a dimensionality reduction technique designed for the visualization of high-dimensional datasets. It proves especially beneficial for datasets featuring a multitude of features, such as images or textual data, where conventional methods may either entail substantial computational costs or lead to the loss of critical information. The fundamental principle underlying t-SNE entails the transformation of high-dimensional data into a lower-dimensional space, all while conserving the intrinsic local structure inherent to the original high-dimensional dataset. The fundamental methodology of t-SNE revolves around the utilization of the k-nearest neighbors (kNN) algorithm to compute the probability distribution for each data point within the lower-dimensional space. This probability distribution hinges upon the distances between data points and their respective kNN counterparts in the high-dimensional space. To ensure that all data points possess a non-zero probability within the lower-dimensional space, t-SNE employs a differentiable variant of kNN, known as perceptron learning.

First, in the original space, t-SNE defines a similarity measure between data points in the high-dimensional space. This similarity is represented by conditional probabilities:

The conditional probability that point j would pick point i as its neighbor if neighbors were picked proportional to their similarity:

$$P_{j|i} = \frac{e^{-\frac{||x_i - x_j||^2}{2\sigma_i^2}}}{\sum_{k \neq j} e^{-\frac{||x_i - x_k||^2}{2\sigma_i^2}}}$$

Here, x_i and x_j are high-dimensional data points, σ_i is the bandwidth for point i , and

$P_{j|i}$ is the conditional probability.

Second, in the lower dimensional space, t-SNE defines a similar conditional probability distribution in the lower-dimensional space. This is done using a t-distribution (Student's t-distribution) with one degree of freedom for the pairwise similarities:

$$Q_{j|i} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq j}(1 + ||y_i - y_k||^2)^{-1}}$$

Here, y_i and y_j are the corresponding points in the lower dimensional space, and $Q_{j|i}$ is the conditional probability in the lower dimensional space.

The goal of t-SNE is to find a lower dimensional representation (y_i) that minimizes the divergence between these two conditional probability distributions. This divergence is typically measured using the Kullback-Leibler (KL) divergence:

$$KL(P||Q) = \sum_i \sum_j P_{j|i} log(\frac{P_{j|i}}{Q_{j|i}})$$

So, t-SNE minimizes the KL divergence between the conditional probability distributions in the high-dimensional space (P) and the lower-dimensional space (Q) by adjusting the positions of y_i in the lower-dimensional space. It does this through an optimization process, usually using gradient descent, to find the y_i values that minimize the KL divergence.

In essence, t-SNE's mathematical formulation is distribution-agnostic; it does not assume that the data follows a certain distribution in the original space. Instead, it focuses on preserving local relationships and pairwise similarities between data points, making it applicable to a wide range of data distributions.

UMAP

UMAP is a dimensionality reduction technique similar to t-SNE but with some differences in its mathematical formulation. UMAP also aims to map high-dimensional data into a lower dimensional space while preserving the underlying structure of the data. Here's a mathematical explanation of UMAP:

First, UMAP starts by constructing a fuzzy simplicial set from the data. The fuzzy simplicial set is represented as an adjacency matrix A, where A_{ij} measures the strength of the connection between data points i and j. This connection strength is determined by considering both the distance between data points in the high-dimensional space and their mutual nearest neighbors.

A_{ij} is computed as follows:

$$A_{ij} = exp(-\frac{d(x_i, x_j) - min(d(x_i, kthneighbour(x_i)), d(x_j, kthneighbour(x_j)))}{\sigma_i \sigma_j})$$

Here, $d(x_i, x_j)$ is the distance between data points i and j, $kthneighbour(x_i)$ is the distance to the k-th nearest neighbour of x_i, and σ_i and σ_j are parameters that control the spread of the fuzzy simplicial set.

Second, After constructing the fuzzy simplicial set, UMAP seeks to find a low-dimensional representation for the data points in a way that preserves the relationships encoded in A. It defines a similar fuzzy set in the lower-dimensional space represented as an adjacency matrix B. B_{ij} is computed as follows:

$$B_{ij} = exp(-\frac{d(y_i, y_j)}{\sigma})$$

Here, y_i and y_j are the low dimensional representations of data points i and j, and σ is a parameter that controls the spread of the fuzzy set in the lower-dimensional space.

UMAP then aims to optimize the low-dimensional representation Y to minimize the KL divergence between the fuzzy simplicial set A and fuzzy set B. This is done by minimizing the following loss function:

$$L(Y) = \sum_i kl_{div}(A_i || B_i)$$

Where $kl_{div}(A_i||B_i)$ is the KL divergence between the i-th row of A and B.

UMAP uses gradient descent to optimize the low-dimensional representations Y to minimize the loss function $L(Y)$. This involves iteratively updating the positions of data points in the lower dimensional space.

In summary, UMAP constructs a fuzzy simplicial set based on the pairwise distances in the high-dimensional space, and it seeks a low-dimensional representation that minimizes the KL divergence between this fuzzy set and a corresponding fuzzy set in the lower-dimensional space. This optimization process results in a meaningful low-dimensional representation of the data.

Chapter 5

Visualization Design

5.1 Preliminary blueprint

For visualization, what researchers are concerned mostly lies in three aspects, which are the origins, the procedures and the results. According to these, the initial structure of the visualization is generated. Figure 5.1 is the sketch of the visualization design.

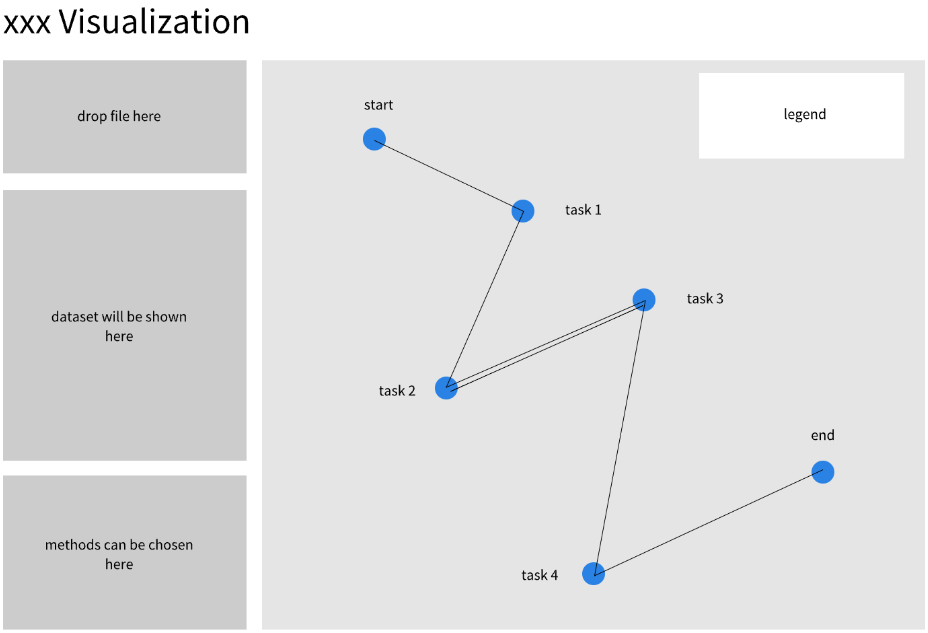


Figure 5.1: Sketch of the Visualization

As is shown in figure 5.1, the components on the left side are related to the origins, which is the related function and options for the datasets.

1. Firstly, there is an area where files can be dragged and dropped, but the format of files may be limited.

2. Next is the data display area, where you can choose which dataset the plot is generated from.

3. Last part is a method selection area for selecting dimensionality reduction methods and color filling rules for the plot generated by the data.

The area next to the left side is the visualization area. The main plot shows the process of sensemaking, where the points represent the actions or tasks in the process and the lines attached two points are on behalf of the flow of the process. Especially, when the user revisits the same webpage or highlights some information, the line will be bolder than others.

On the right side is the Legend of the plot. It will show the color representation of the plot.

5.2 Final Version

With deeper research on the project and exploratory analysis of data, some ideas came to my mind, such as showing more details of the data. By discussing these ideas with supervisor, I gradually improved my sketch and formed the final version of the visualization webpage. The final version of webpage looks like as follows:

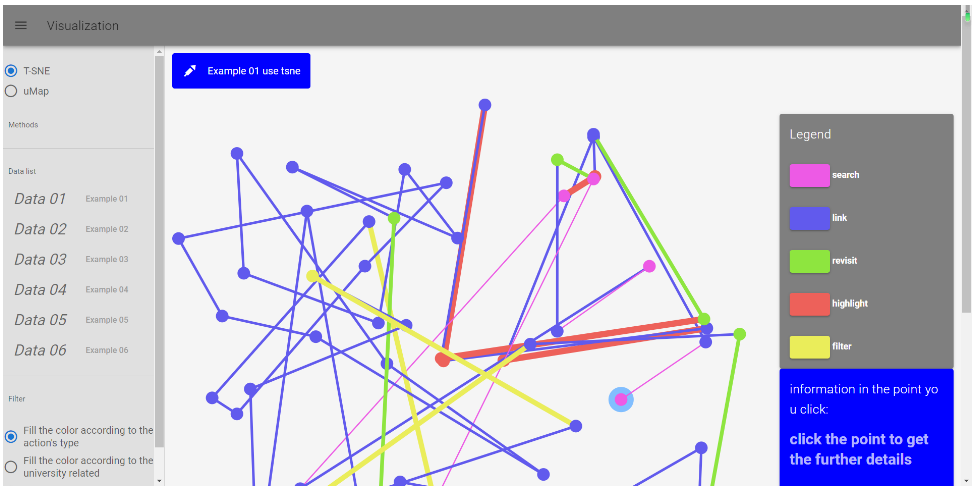
In figure 5.2, most parts are as expected before, while there are some changes:

- I remove the file area, the reasons are discussed in 6.2.

- After choosing the dataset, the results, as well as the key words, will be shown on the right side of the plot.

- The method selection area is divided into two parts, methods area and filter area.

With these changes, it will be more convenient to use and adapt to various situations. Details will be demonstrated in 7



● T-SNE
○ uMap

● Fill the color according to the action's type
○ Fill the color according to the university related

information in the point yo u click:

click the point to get the further details

Figure 5.2: The generated plot

Chapter 6

Implementation

As mentioned in the previous Chapter, the aim of the project is to develop a visualization tool to analyze the user's behavior. Therefore, in this Chapter, we will interpret how to apply the methodology to practice, combining algorithms with visualization, and utilize this visualization tool in further analysis.

6.1 Theoretical definition

For clear classified purposes, I simply divide the provenance into three layers according to the richness in its contents.

1. **Goal/ Task**: (highest level) the goal of the process or the task of sensemaking such as "the top 3 universities in Europe".

2. **Action/ Sub-task**: (medium level) the action of the user and the sub-task in each task such as "highlight the name of university".

3. **Event**: (lowest level) the interaction event of a user such as "click the point".

Based on the theory, we can give each group of data a type of action in order to sort the action according to its importance as mentioned in 3.2.1
The task in the project is to choose 3 universities as a result (recorded in results.json). Additionally, there are four types of action that are related to the first, second, third or none of them.

6.2 Data processing

In order to obtain the data we need, after collection by SensePath, the data needs some manual processing before conversion to vector sequence.

1. When Python loads the .json files, it is necessary to remove the beginning structure of .json in case that the file cannot be read by Python.

2. Before we process the main data, we should delete some redundant, duplicate, and unnecessary data. For example, ”id” is the data only for counting, so that it needs to be simplified as a small number to reduce the possibility of index errors caused by it.

3. Some incorrect data may be collected, and we also need to remove them or modify them manually according to the correct related data.

6.3 How to do the vectorization in each action

In order to obtain the meaning of the text in each action and convert it into a vector for visualization, this project chooses to use the embedding layer of the LangChain model as the data processing method of conversion to vectors. LangChain.embeddings performs well in vectorization, especially in short texts. More over, it is based on neural translation system, so it runs faster.

6.4 How to recognize the flow of information

As mentioned before, one of the aim of visualization tool is to help the researchers recognize the flow of information. For the provenance data, the timestamp can be regarded as a criterion for determining the order between actions, while the calculated duration can be used for the importance of certain information, which will probably be used for further research.

6.5 Data visualization and interaction

To facilitate user interaction with visualization tools, the imperative task is the creation of a web page that accommodates these functions. Primarily, this entails front-end development, with HTML and JavaScript being the primary programming languages of focus. As previously noted, there exist two prominent front-end frameworks for our consideration: Vue and React. After setting up the environment and some debugging, we ultimately chose Vue. The main reason is that although React responds quickly, there are some problems in the local computer environment configuration. In limited time, we did not find a suitable solution, so we chose Vue, which is characterized by simplicity and lightness. It is built upon standard HTML, CSS, and JavaScript, providing a declarative, component-based programming model to assist developers in the efficient creation of user interfaces. During the development process, standard CSS alone may not suffice to meet the project's requirements comprehensively. Hence, for more efficient and aesthetically pleasing layout and components, here we will use a relevant virtual component framework called Vuetify. Vuetify's collection of UI components maintains a consistent style throughout the entire application and offers ample customization options to cater to any use case.

In the initial design in Chapter 5, there is an area for user to drop file. However, in 6.2, we have explained that some data needs to be manually processed, thus, directly importing data is not appropriate. Furthermore, owing to the asynchronous characteristics inherent in file reading, this capability was omitted in the final version. Presently, users are restricted to observing the visualizations of 6 processed examples.

Chapter 7

Outcomes and Future

7.1 Results and Analysis

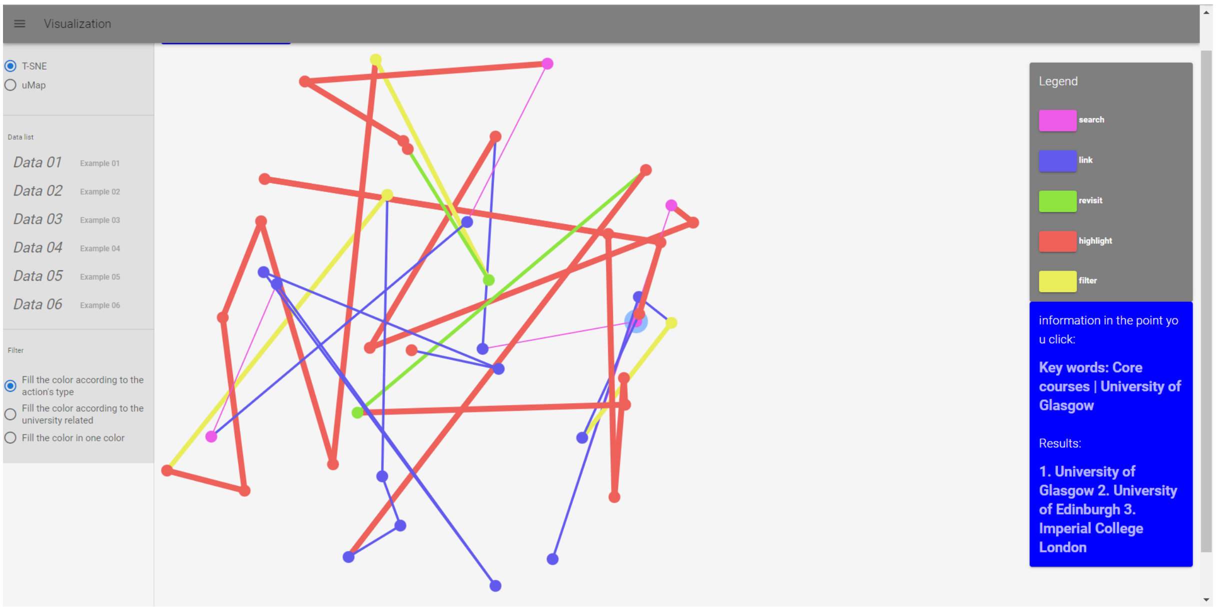
In this project, the aim is to help researchers obtain the hidden insight and the flow of information from the data, which may facilitate further research.

Throughout the visualization tool developed in this project, the researchers can easily find out what the task is in the process, how the decisions are made and what supports the results. The results and analysis of the visualization tool are as follows:

- In figure 7.1, this is a situation that the results are known in advance, which are displayed at the bottom of the panel. A point on one end of a red line has been selected, and the information of this point is shown in the right panel, which says 'key words: core courses — University of Glasgow'.

  As the results are known in advance, we can figure out that the highlight actions are all related to the results, while some of the revisit actions are related to the results and the others are less related to the results.

  In this situation, we are more concerned about the flow of information, the process of decision making and the understanding of provenance data. And these are precisely



● T-SNE
○ uMap

● Fill the color according to the action's type
○ Fill the color according to the university related
○ Fill the color in one color

information in the point you click:

Key words: Core courses | University of Glasgow

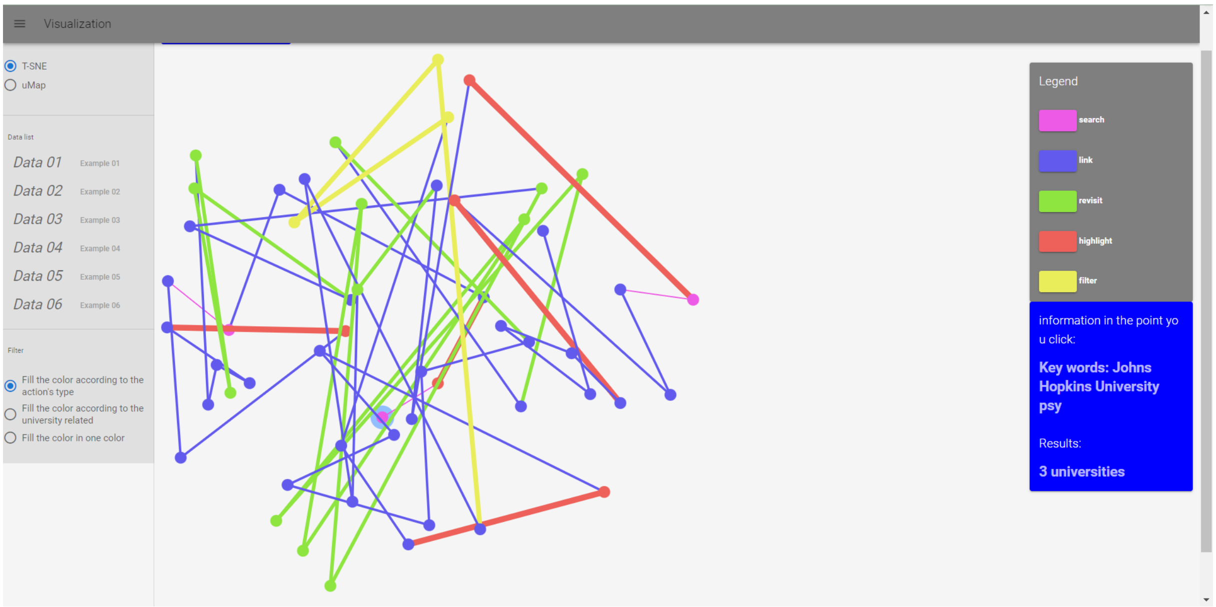
Results:

1. University of Glasgow 2. University of Edinburgh 3. Imperial College London

Figure 7.1: Situation 1: known results

filtered by this visualization tool.

- In figure 7.2, this is another situation that the results are hidden in advance.



● T-SNE
○ uMap

● Fill the color according to the action's type
○ Fill the color according to the university related
○ Fill the color in one color

information in the point you click:

Key words: Johns Hopkins University psy

Results:

3 universities

Figure 7.2: Situation 2: hidden results

Unlike situation 1, results are not displayed at the bottom of the panel, so we need to conduct some analysis and statistics on the plot. Certainly, the focus is still on the points at both ends of the red line representing the highlight actions and the green line representing the revisit actions. After getting insight of the points, we collect that the actions related to University of Illinois– Urbana-Champaign are 10 in all, the ones related to Duke University are 3, and the ones related to John Hopkins University are 2. Therefore, we draw a conclusion that the top 3 universities in this process are 1. University of Illinois– Urbana-Champaign, 2. Duke University, and 3. John Hopkins University. Compared with the hidden results, it is completely consistent.

While under this condition, we pay more attention to the analysis of tasks, exploring the hidden insight of the data and obtaining some data that may be helpful for future research. Since that, The achievement of these goals depends on the proper use of this visualization tool.

7.2 Outcomes

The "Projection of Provenance Vector Sequence" project represents a significant endeavor in the domain of data lineage and information system analysis. After research in relevant fields and the development of visualization tools, this project has achieved milestones in understanding of data provenance and transformation of history data.

As for the author of this article, I have learned a lot from this project, including some relatively cutting-edge methods and technologies in this field, as well as research on some new models, which has greatly benefited me in future research and work.

For the researchers who use this tool, as the tool leveraged a combination of topology-driven and attribute-driven layouts, offering users distinct perspectives on data provenance, researchers can use it to meet the analysis needs of different purposes. The introduction of novel strategies for vectorization and quantization of data provenance enriched the depth of insights that could be gleaned from the visualizations.

7.3 Challenge and Future

This project presents certain challenges. These include handling large-scale and dynamic datasets, ensuring scalability and interactivity in visualization techniques, and addressing user-centric design considerations. Future research should focus on developing novel visualization methods that effectively address these challenges, integrate machine learning and AI techniques, and incorporate user feedback and evaluation methodologies, offering cloud-compatible solutions for evolving data environments, fostering community collaboration for diverse perspectives, and maintaining a relentless focus on usability and accessibility through regular user feedback and studies. These directions promise to enrich the project's capabilities and make it even more valuable to researchers and professionals in the field of data lineage and information system analysis. As the learning capability of machine learning and AI techniques improves and becomes increasingly mature, The outcomes of this project will also be more widely used in various fields.

Chapter 8

User Evaluation

8.1 Aim

In order to gain a detailed understanding of the advantages and disadvantages of this visualization tool, four students were invited to use this tool and asked to express their experiences.

8.2 Evaluation Summary

There are aspects on evaluation:

- **Function**: Evaluate the condition of basic functions of the visualization tool
- **Usability**: Evaluate the difficulties of interaction between visualization tool and users
- **Appearance**: Evaluate the appearance of visualization tools, including the webpage layout, color, and display
- **Task Performance**: Assess how well users can perform in exploring and analyzing data provenance using the visualization tool.

The table 8.1 shows the evaluation result:

The table gives us a direction for further optimization.

	Advantages	**Disadvantages**
Function	all went well	simple
Usability	smooth using	need to choose all when changing
Appearance	color and webpage layout are good	Image not displaying all
Task Performance	performed well	better if prediction added

Table 8.1: User evaluation summary

Chapter 9

Summary and Reflections

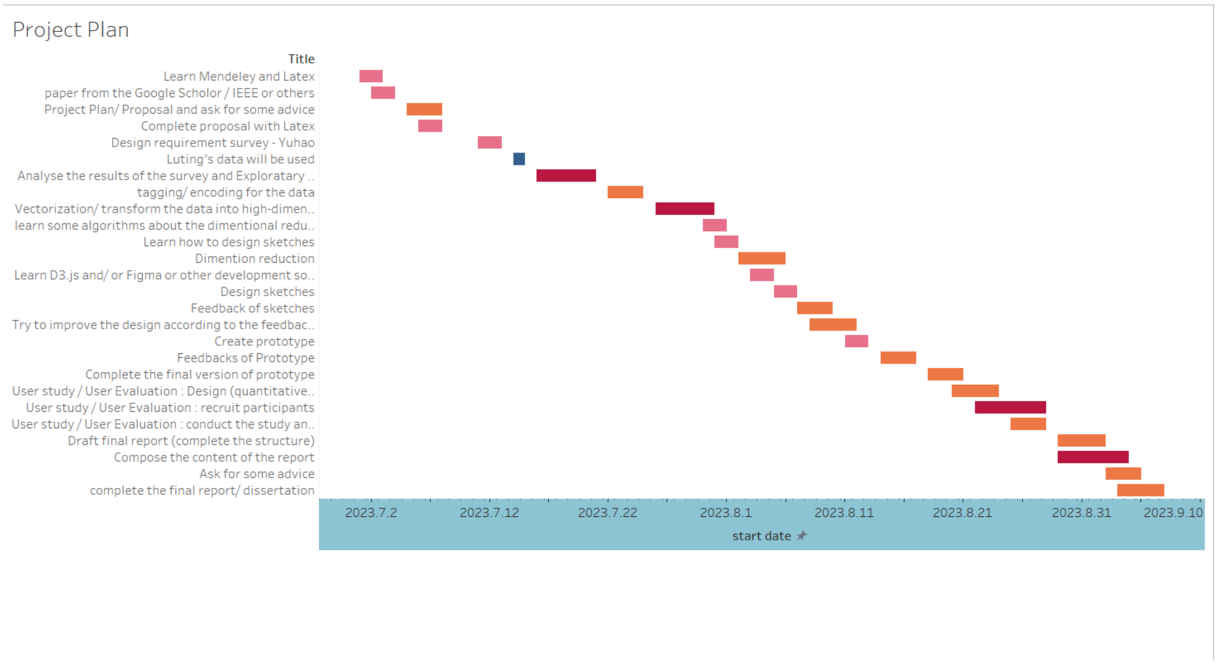


Figure 9.1: Gantt chart of project

9.1 Project management

Most project management is done on Github, with the link at https://github.com/Vis4Sense/student-projects/tree/main/2023-summer/yu-hao. Meanwhile, the code and related supplementary documents are also included.

9.2 Project plan

The entire project can be done mainly due to the understanding and relevant knowledge of this project, while programming is relatively not that difficult. From the beginning of the project, studying related knowledge as much as possible and learning the required skills are the key to the project, so that a direction of the project will be found. After that, it is best to execute as per the planned schedule shown in figure 9.1. If there are any questions about the project, they can be discussed with the supervisor during the weekly meetings and most of the problems can be settled. Similarly, if there are any new ideas, they can be shared with the supervisor during the weekly meetings and you will be given some advice. The management and progress of the project are also managed through Github and the meetings with the supervisor and teammates.

9.3 Actual completion

Most of the tasks in the plan are completed as expected, but due to some personal reasons, some tasks were not completed on time according to the Gantt chart. But as for the results, the whole project was successfully completed. Due to the tight schedule of the project, some ideas were not implemented in time, such as finding more models for vectorization. There are also some comparative tasks that have not been done in time, such as comparing the differences in visualization tools developed by React and Vue, and so on

9.4 Contributions and reflections

The project lasts about three months. In the first month, the author has been busy reading literature and learning relevant knowledge in the field, and completed the proposal. In the second month, the author started to collect data. Since SensePath was removed from the Chrome store, the author had no way to collect new data. After discussing this with the supervisor, the author decided to use the data collected by the predecessors and did relevant processing. In the last month, the author mainly focused on the development of visualization tools and the writing of the final report. The final outcome of the project is shown as figure 5.2.

Bibliography

[1] Alharthi, S. A., LaLone, N. J., and Sharma, H. N. An activity theory analysis of search and rescue collective sensemaking and planning practices. *Conference on Human Factors in Computing Systems - Proceedings* (5 2021).

[2] de Sousa, M., and Gonçalves, A. humanportal – a react.js case study. In *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)* (June 2020), pp. 1–6.

[3] Dettmers, T., Lewis, M., Belkada, Y., and Zettlemoyer, L. Llm.int8(): 8-bit matrix multiplication for transformers at scale, 2022.

[4] Duan, Y., Liu, C., Li, S., Guo, X., and Yang, C. An automatic affinity propagation clustering based on improved equilibrium optimizer and t-sne for high-dimensional data. *Information Sciences 623* (2023), 434–454.

[5] Endert, A., Ribarsky, W., Turkay, C., Wong, B. L., Nabney, I., Blanco, I. D., and Rossi, F. The state of the art in integrating machine learning into visual analytics. *Computer Graphics Forum 36* (12 2017), 458–486.

[6] Henry, W. C., and Peterson, G. L. Sensorre: Provenance support for software reverse engineers. *Computers Security 95* (8 2020), 101865.

[7] im Studiengang Bachelor, Prüferin, B., Zweitgutachter, U. S., geb Knoblauch, M. B., and Wohlgethan, E. Bachelorarbeit eingereicht im rahmen der bachelorprüfung.

[8] Jia, K., Meng, F., Liang, J., and Gong, P. Text sentiment analysis based on bert-cblbga.

[9] Kobak, D., and Linderman, G. C. Initialization is critical for preserving global data structure in both t-sne and umap. *Nature Biotechnology 2021 39:2 39* (2 2021), 156–157.

[10] Maaten, L. V. D., and Hinton, G. Visualizing data using t-sne. *Journal of Machine Learning Research 9* (2008), 2579–2605.

[11] Maltsev, A. S., Umarova, N. N., Pashkova, G. V., Mukhamedova, M. M., Shergin, D. L., Panchuk, V. V., Kirsanov, D. O., and Demonterova, E. I. Combination of total-reflection x-ray fluorescence method and chemometric techniques for provenance study of archaeological ceramics. *Molecules 28*, 3 (2023).

[12] McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

[13] Min, J. W. W., Jali, N., and Chai, W. Y. Vehicle tracking system, 2021.

[14] Nguyen, P. H., Xu, K., Wheat, A., Wong, B. W., Attfield, S., and Fields, B. Sensepath: Understanding the sensemaking process through analytic provenance. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2016), 41–50.

[15] Pokorny, V. J., Sponheim, S. R., and Rawls, E. Impact of reduced-dimensionality independent components analysis on event-related potential measurements. *Psychophysiology 60*, 5 (2023), e14223. e14223 PsyP-2021-0516.R1.

[16] Pujara, M. N. Industrial internship report building a dynamic car dealership website with react js.

[17] Ramos, G., Suh, J., Ng, R., and Meek, C. Forsense: Accelerating online research through sensemaking integration and machine research support.

[18] Serna-Serna, W., de Bodt, C., Alvarez-Meza, A. M., Lee, J. A., Verleysen, M., and Orozco-Gutierrez, A. A. Semi-supervised t-sne with multi-scale neighborhood preservation. *Neurocomputing 550* (2023), 126496.

[19] Shaheer, S., Hossain, I., Sarna, S. N., Kabir Mehedi, M. H., and Rasel, A. A. Evaluating question generation models using qa systems and semantic textual similarity. In *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)* (March 2023), pp. 0431–0435.

[20] Song, J., Zhang, M., and Xie, H. Design and implementation of a vue.js-based college teaching system. 59.

[21] Thakkar, M. *Introducing React.js.* Apress, Berkeley, CA, 2020, pp. 41–91.

[22] Topsakal, O., and Akinci, T. C. Creating large language model applications utilizing langchain: A primer on developing llm apps fast. *International Conference on Applied Engineering and Natural Sciences 1*, 1 (Jul. 2023), 1050–1056.

[23] Ujas, T. A., Obregon-Perko, V., and Stowe, A. M. *A Guide on Analyzing Flow Cytometry Data Using Clustering Methods and Nonlinear Dimensionality Reduction (tSNE or UMAP).* Springer US, New York, NY, 2023, pp. 231–249.

[24] Walchshofer, C., Hinterreiter, A., Xu, K., Stitz, H., and Streit, M. Provectories: Embedding-based analysis of interaction provenance data. *IEEE Transactions on Visualization and Computer Graphics* (2021), 1–1.

[25] Xu, K., Ottley, A., Walchshofer, C., Streit, M., Chang, R., and Wenskovitch, J. Survey on the analysis of user interactions and visualization provenance. *Computer Graphics Forum 39* (6 2020), 757–783.

[26] You, E. Vue. js-the progressive javascript framework— vue. js, 2014.