README: A Literature Survey Assistant

Raghav Goyal, Shih-Han Chou, and Siddhesh Khandelwal
{rgoyal14, shchou75, skhandel}@cs.ubc.ca

Abstract—Literature review is an integral element of academic research, enabling researchers to learn about and build on existing work. Traditionally, this involves manually going through various published articles, either through following the citations in a reference paper, or via keywords on sites like Google Scholar. This process can often be tedious, and there is a high likelihood of missing out on certain related literature, owing to the sheer volume of publications every year. In addition, analyzing the advances and progression in a field requires a holistic view, which manual iteration over papers lacks. To this end, we propose README, an interactive tool aimed to aide with literature reviews. README would not only enables users to obtain a holistic view of various papers and topics, but also identifies and recommends relevant papers, given a reference paper the user is interested in. Additionally, it allows for chronological sorting of applicable papers, thus making analysis of trends and patterns much easier.

1 INTRODUCTION

Literature reviews are crucial components to sustainable academic research. The knowledge of existing research provides a stable foundation to build upon, while simultaneously avoiding redundancy, facilitating improvements, and enabling meaningful contributions. Despite its importance, by far the most common way to conduct literature search is to manually sieve through large collections of papers, one at a time, to find work that is relevant. This exploration is done either via hopping through references within a paper, or using keywords to retrieve information from engines like Google Scholar. Due to the plethora of published work, this approach is often extremely time intensive and increases the possibility of missing out on relevant research.

We argue for the need of a specialized framework that allows researchers to only focus on a few pertinent publications, while simultaneously providing a aggregated overview of the topic diversity in a particular field at a glance, saving both time and effort. To this end, we aim to develop README: an interactive literature review tool aimed at making surveying relevant research papers easier. This proposed prototype would serve three distinct purposes:

P1: Provide a holistic view of publications in a particular area clustered by sub-topics, wherein the clustering is inferred via two dimension projections of text-based vector representations for each paper.

P2: For a particular query paper, recommend a small number of publications using a similarity and topic-coverage based relevance algorithm.

P3: Conditioned on a seed paper, aide the analysis of publication trends in the relevant domain via a chronological view.

The tool is intended to be interactive, enabling seamless transition between the aforementioned three use-cases. Figure 1 provides an overview of the proposed framework.

This proposal is structured as follows: We first list the hypotheses assumed in order to achieve the aforementioned purposes, then we describe the dataset, preprocessing and and various abstractions required, and finally we discuss the elements of our proposed visualization.

2 HYPOTHESES

Here we describe certain hypotheses we assume to hold in order efficiently implement the proposed tool. Through this project, we also aim to validate these hypotheses and hopefully learn about some salient characteristics of the task that might be helpful in the future.

H1: Each paper \( p \) can be converted into a high dimensional vector embedding \( e_p \), wherein the embedding is expressive enough to capture the “essence” of the corresponding paper. That is, two similar papers...
Based on biological control strategy in pest management, we construct and investigate a pest-epidemic model with impulsive control, i.e., periodic spraying microbial pesticide and releasing infected pests at different fixed moments. By using Floquet theorem and comparison theorem, we prove that the pest-eradication periodic solution is globally asymptotically stable when the impulsive period ...
High-dimensional InfoVis solution Embedding visualization Abstraction we currently have in mind is using the Embedding space similarity where \( \alpha \) where

A key objective of README is to recommend a set of relevant publications given the user is interested in a particular paper. As described in Section 2, \( \mathbf{H3} \) provides two definitions of relevance: \( i) \) similarity in the high dimensional embedding space, and \( ii) \) coverage in the topic space. We now discuss the details regarding these two criterias.

**Embedding Space Similarity:** Given two papers \( p \) and \( q \), the similarity \( s_{pq} \) between them can be computed using an inverse cosine distance between their corresponding embeddings \( \mathbf{e}_p \) and \( \mathbf{e}_q \). Specifically,

\[
s_{pq} = \frac{\mathbf{e}_p \odot \mathbf{e}_q}{||\mathbf{e}_p|| \cdot ||\mathbf{e}_q||}
\]

where \( \odot \) is the inner product. We conducted some preliminary analysis using this similarity measure. Figure 3 highlights an example showing the most similar titles/abstracts to a particular query paper.

**Topic Space Coverage:** As described in Section 4.2, let \( \alpha_p = [\alpha_1^p, \ldots, \alpha_k^p] \) be the topic scores computed for a query paper \( q \). For each \( \alpha_i^p > \tau \), where \( \tau \) is a positive threshold, our aim then is to find a publication \( p_i \) that: \( i) \) has a high \( \alpha_i^p \) value indicating that topic \( i \) is present in paper \( p_i \), and \( ii) \) \( p_i \) and \( q \) are really similar. We still need to explore what the ideal “similarity” metric would be. One option we currently have in mind is using the Embedding space similarity described in Section 4.3. Intuitively, our goal is to present the user with a set of papers \( \mathbf{P} = \{p_1, \ldots, p_k\} \) that collectively spans the same set of topics as the query \( q \).

To provide a holistic view of the papers clustered by sub-topics, we first project the high dimensional embeddings \( \mathbf{e}_p \) computed in Section 4.1 2-D space using t-SNE [7]. To avoid visual cluttering, these 2-D points can then be grouped into different topics and visualized as a scatter plot, as shown in Figure 5(a). Following hypothesis \( \mathbf{H1} \), the hope here is that similar papers would be closer together in the projected 2-D space. The user can explore further into a particular topic by clicking on the appropriate cluster, which will provide a zoomed-in view as shown in Figure 5(b). Each small circle in Figure 5(b) is a paper.

The user can also select a specific query paper, either from a list obtained via a key-word search, or selecting a circle in the holistic view. For this query paper, our aim then is to recommend a set of relevant papers, using the two relevance measures defined in Section 4.3. Our initial design involves visualizing these relevant papers in a star-shaped structure, as shown in Figure 5(c). Here the selected paper is highlighted in red. The user will also have the option to choose between the relevance measures, each providing a different set of relevant papers. The thickness of arrows in 5(c) indicate the relevance strength, the solid lines correspond to the relevant papers that were references in the query paper, and the dashed lines are papers that are relevant but are not cited by the query paper.

To aide with trend analysis, for a selected query paper, the user can also obtain a chronological view of papers. As shown in Figure 5(d), each node is a paper, and their position on the horizontal axis corresponds to the time when that paper was published. To avoid visual clutter, we only show a handful of papers at each depth. These papers are selected by recursively applying our relevance algorithm on each node.

We detail the steps and our findings during our implementation:

**Filter data.** We took 10^8 version of DBLP Citation Network dataset which contains about 3 mil papers. It's cumbersome to perform any kind of analysis at such a large scale dataset, therefore we filter papers by conference venues. Since we are more accustomed to the field of machine learning, deep learning and its applications, we selected venues by keywords: "machine learning", "computer vision", "artificial intelligence", "machine translation", "robotics", "language processing", "image understanding". We obtained a total number of 156,857 papers
**Fig. 5: Proposed InfoVis solution**. (a) Grouped topics, (b) Zoomed-in View, (c) Recommend set of papers, (d) Chronological view for trend analysis.

**Fig. 6: Dimensionality reduction using paper titles v/s abstracts**. The projections given by t-SNE is more informative and discriminative when learned on paper abstracts (b) as compared to their titles (a).

**High-dimensional embedding.** Each paper $p$ can be represented as a high dimensional embedding $e_p$. We leverage titles and abstracts of papers to obtain their embeddings. We use SciBERT [2] model which is pre-trained on the corpus of scientific text semanticscholar.org. We encode both the titles and abstract of the papers in our dataset. For encoding, we use the encoder Transformer [15] module of SciBERT which gives a 768-dimensional embedding for each token. We then average out the embeddings of all tokens in a title or abstract to obtain the final embedding.

**t-SNE on titles v/s abstract.** Currently, we filter out papers by conferences otherwise we end up with a hair ball of points in their 2D t-SNE plot. We experimented with three conferences on different areas: EMNLP (language), BMVC (vision), and RAS (robotics); each of them contains 1697, 1345, and 2124 paper respectively. As expected, the representations are more informative when the abstracts are encoded (Fig. 6b) as compared to titles (Fig. 6a).

5.1 Case study of a paper

Here we present a walk-through of our approach when a paper is selected along with the visualization.

**Vis of all data.** We picked “Computer Vision and Pattern Recognition (CVPR)” conference containing 9,723 papers which when combined with their references, comes up to 42,242 papers. Performing t-SNE dimensionality reduction yields the plot shown in Fig. 7a.

**Vis of selected paper w/ direct references.** We randomly selected a paper titled “Efficient Action Localization with Approximately Normalized Fisher Vectors”, and plotted the paper and its direct references shown in the Fig 7b.

**Vis of selected paper w/ top-100 closest papers.** We computed Euclidean distance of the selected paper with all the other 9,723 CVPR papers in 768-dimensional space and picked 100 closest papers to it shown in Fig. 7c.

**Vis of selected paper w/ top 3 topics.** We use the probabilistic topic modeling approach called Latent Dirichlet Allocation (LDA) [3], and extracted 50 topics from the dataset consisting of abstract of 9,723 CVPR papers. For the selected paper, we compute its topic distribution and filter top 3 most probable topics. For each of those topics, we pick three most likely paper from the list of top-100.
(a) t-SNE plot on abstracts of CVPR conference papers along with their references, with 42,242 as total number of points.

(b) t-SNE plot with selected paper and its direct references.

(c) t-SNE plot with selected paper and its top-100 closest papers in 768-dimensional space.

(d) t-SNE plot with selected paper with its three most probable topic, where each topic has three most likely paper from the top-100 closest papers to selected paper.

Fig. 7: A placeholder visualization of case study. We start with plotting all papers (a), then selecting a paper and plotting it along with its references (b). We then sort and pick 100 closest papers to the selected paper in high-dimensional space and plot them (c). And finally, we use the topics extracted using LDA [3] to select three topmost topics and present our recommendation of papers within each topic (d).

5.2 Alternate Infovis Solutions
We considered alternate Infovis solutions to this problem before converging on the one shown in Figure 5. Here we detail other ideas, and discuss possible reasons as to why we thought they wouldn’t be suitable for this task.

Scatterplot With Link Marks: As the dataset is in the form of a citation network, our initial thought was to visualize all the papers using a point mark. An additional interactive interface that would then allow the user to select a particular paper, which would in-turn superimpose line marks linking the selected paper to the most relevant subset of recommended paper. Similar to Section 4.1, this would involve deriving a high dimensional representation from each paper, and then reducing this dense representation to a two dimensional feature space. These 2D points could then be visualized as a scatterplot. The immediate downside of this approach is the extremely large number of papers in the dataset (∼3 million), which makes using a naive scatterplot infeasible and visually cluttered (Figure 7a shows a subset of the total papers, which is still hard to interpret).

Graph with Link Marks: To solve the aforementioned issue, our next approach was to require the user to provide a query or seed paper. This query paper would be the basis for the recommendation, thus removing cognitive overload when looking at all papers in a scatterplot. Borrowing inspiration from [8], the query paper would be represented as a point mark, with links going in the upwards and downwards direction, connecting the query paper to relevant papers. This process would be recursively applied, where the relevant papers would be further connected to their relevant paper. The Y-axis would encode time, where the most recent paper would be on the top. This approach, however, doesn’t provide the user with a general overview of the field and limits exploration as a seed paper is required as input. In addition, the recursive graph generation could lead to visual cluttering as the size of the graph increases exponentially with each level.
Our proposed solution in Figure 5 tries to address the issues of the two alternate solutions. The grouped scatterplot (Figure 5(a,b)) allows the user to view broad groups and traverse through them without being overloaded with information. The recommender view (Figure 5(c,d)) suggests relevant papers based on a selected query (which could come from exploration of the scatterplot), and allows for different views and selection criterias to further aid the user in research.

5.3 Proposed Solution Limitations

Our proposed solution has its own set of limitations. For one, it still doesn’t scale well to millions of papers, as even the aggregated scatterplot gets too crowded with more number of topics and conferences. Additionally, our relevance algorithm relies on topic/similarity heuristics, and there might be situations where it misses out on certain relevant recommendations. Even though we expect it to perform reasonably well, more experimentation and analysis is required to have a better understanding.

6 Related Work

From visualization perspective, a 2D scatter plot of papers formed using a dimensionality reduction technique such as t-SNE [7] or UMAP [9] is not new and has been used extensively to visualize related papers where closeness in 2D space represents similarity [1, 6]. ICLR’s Paper Explorer [1] uses an interactive scatter plot where hovering over a paper’s point mark produces its title, author and a representative figure. Adjutant [6] goes a step further to form topic clusters from related papers in an unsupervised fashion that allows for topic-based exploration. In this work, we will adopt the similar workflow, but plan to visualize not only the related papers, but also their chronology to investigate evolution of topics and more broadly the research areas.

For topology of chronological ordering of related papers, we plan to take inspiration from Overview [5], where they proposed a tree-based visualization of hierarchically clustered documents intended for search and exploration.

A related CPSC 547 course project PaperQuest [12] proposes a multi-level filtering of relevant papers together with user’s interest and preference. However, in this work we consider a single-level decision based on user’s query only.

7 MILESTONES

Our plan for implementing the tool is as follows:

- High dimensional embedding: Instead of sticking on the current pre-trained language model, we plan to try different language embeddings such as GloVe [11] or Word2Vec [10] embeddings to embed the papers to high dimensional space.

- Topic modeling: Accurately training a topic model is crucial to the performance of our topic-based relevance measure. We aim to look at different variants of the model, and also try training it on different inputs (like abstracts, titles, or both).

- Infovis solution: Our initial thought is to use the D3 [4] framework to implement our visualizations. As none of use are proficient with the framework, we will explore other possibilities as well.

Below is a rough plan for how we plan to split the work among group members. We plan to finish the feature embedding and topic modeling portions of the project separately, and then ideate and implement on the InfoVis portion together. Table 2 provides a detailed breakdown of the milestones and hour estimates.

Table 2: Milestones. Precise milestone specifications and hour estimates. RG: Raghav Goyal, SHC: Shih-Han Chou, SK: Siddhesh Khandelwal.

<table>
<thead>
<tr>
<th>Task</th>
<th>Est. hours</th>
<th>Actual hours</th>
<th>Est. completed date</th>
<th>Actual completed date</th>
<th>Workload distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data gather and filtering</td>
<td>10</td>
<td>12</td>
<td>Oct 24</td>
<td>Oct 26</td>
<td>RG: Raghav Goyal, SHC: Shih-Han Chou, SK: Siddhesh Khandelwal</td>
</tr>
<tr>
<td>High dimensional embedding</td>
<td>6</td>
<td>8</td>
<td>Nov 7</td>
<td>Nov 10</td>
<td>RG: Raghav Goyal, SHC: Shih-Han Chou, SK: Siddhesh Khandelwal</td>
</tr>
<tr>
<td>Topic modeling</td>
<td>10</td>
<td>15</td>
<td>Nov 14</td>
<td>Nov 14</td>
<td>RG: Raghav Goyal, SHC: Shih-Han Chou, SK: Siddhesh Khandelwal</td>
</tr>
<tr>
<td>Placeholder Infovis solution</td>
<td>10</td>
<td>8</td>
<td>Nov 16</td>
<td>Nov 17</td>
<td>RG: Raghav Goyal, SHC: Shih-Han Chou, SK: Siddhesh Khandelwal</td>
</tr>
<tr>
<td>Proposed interactive Infovis solution</td>
<td>25</td>
<td></td>
<td>Dec 5</td>
<td></td>
<td>RG: Raghav Goyal, SHC: Shih-Han Chou, SK: Siddhesh Khandelwal</td>
</tr>
</tbody>
</table>

Table 1: Recommended papers by topics in the case study. For a selected paper three most probable topics are chosen extracted by LDA [3]. For each topic three paper recommendations are made from the set of top 100 closest papers to the selected paper.

<table>
<thead>
<tr>
<th>Task</th>
<th>Selected paper</th>
<th>Paper Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1 [‘action recognition’, ‘multi class’]</td>
<td>Recommended paper 1</td>
<td>Efficient feature extraction, encoding and classification for action recognition</td>
</tr>
<tr>
<td></td>
<td>Recommended paper 2</td>
<td>Optimizing 1-Nearest Prototype Classifiers</td>
</tr>
<tr>
<td></td>
<td>Recommended paper 3</td>
<td>Beyond Gaussian Pyramid: Multi-skip Feature Stacking for action recognition</td>
</tr>
<tr>
<td>Topic 2 [‘high dimensional’, ‘image features’]</td>
<td>Recommended paper 1</td>
<td>Lost in quantization: Improving particular object retrieval in large scale image databases</td>
</tr>
<tr>
<td></td>
<td>Recommended paper 2</td>
<td>Bundling features for large scale partial-duplicate web image search</td>
</tr>
<tr>
<td></td>
<td>Recommended paper 3</td>
<td>Learning Separable Filters</td>
</tr>
<tr>
<td>Topic 3 [‘widely used’, ‘recognition performance’]</td>
<td>Recommended paper 1</td>
<td>Chebyshev approximations to the histogram χ 2 kernel</td>
</tr>
<tr>
<td></td>
<td>Recommended paper 2</td>
<td>Grassmann Averages for Scalable Robust PCA</td>
</tr>
<tr>
<td></td>
<td>Recommended paper 3</td>
<td>Effective learning-based illuminant estimation using simple features</td>
</tr>
</tbody>
</table>


