README: A Literature Survey Assistant

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

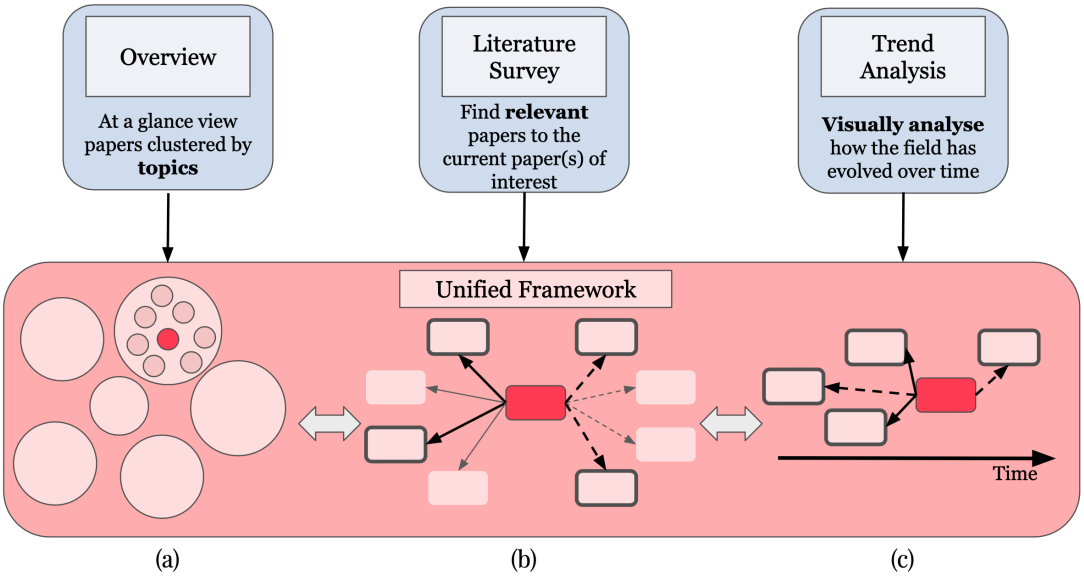

Fig. 1. **Task setup**. Our proposed framework serves three purposes. (a) Overview of publications clustered by topics, (b) Relevant recommendations conditioned on a paper of interest, and (c) Trend analysis via a chronological view.

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 $\mathbf{e}_p$, wherein the embedding is expressive enough to capture the "essence" of the corresponding paper. That is, two similar papers

Abstract	Based on biological control strategy in pest management, we construct and investigate a pest-epidemic model with impulsive control, i.e., periodic spraying microbial esticide 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 ….	Agitation in a mixer-settler is one of the most common operations, yet presents one of the greatest challenges in the area of computer simulation. Mixer-settlers typically contain an impeller mounted on a shaft, and optionally can contain baffles. The hydrodynamic characteristics of mixer-settlers have been studied in the present study. The effect of ….
Authors	Guoping Pang, Lansun Chen	Mohsen Ostad Shabani, Mehdi Alizadeh, Ali Mazahery
Paper id	4aa69add-3978-480b-a1c0-d99a83d7e324	4aa6bba8-ff79-49e2-b42c-d9ac62f512f2
#Citation	8	0
References	04754a28-6bf4-4d5d-8e42-2677d8564cdc, 33a877a6-9f28-4762-bf48-81439abfd19b, 7970c424-ac2b-43d5-9e67-07b07f15bfa5, a820c9c1-87ce-4241-8706-eca4aa9094e8, c1008c92-dcdb-444a-8018-f20a692b02b5	d3ac318b-2873-4583-aede-bed9e89d82dd
Title	Dynamic analysis of a pest-epidemic model with impulsive control	Fluid flow characterization of liquid–liquid mixing in mixer-settler
Venue	Mathematics and Computers in Simulation	Engineering With Computers
Year	2008	2011

Fig. 2. Examples of DBLP Citation Network Dataset. For each data, it contains the abstract of the paper, authors, paper id, number of citations, references, venue, and published year. Due to the space limitation, we only show a portion of the abstract.

Selected Paper (Title)	References (Title)	Similarity
Automated regression testing using DBT and Sleuth	TestTube: a system for selective regression testing	0.73
	Selective regression testing—assumptions and fault detecting ability	0.76
	A framework for evaluating regression test selection techniques	0.74

Selected Paper (Abstract)	References (Abstract)	Similarity
Regression testing is an important activity in software maintenance. Current regression testing strategies can be categorized into two groups: 'retest all' and 'selective regression' testing. Each of these two groups encompasses a variety of strategies. In industrial practice, regression testing procedures vary widely. Sometimes, several regression testing techniques are used in combination. Technique selection …	The paper describes a system called TESTTUBE that combines static and dynamic analysis to perform selective retesting of software systems written in C. TESTTUBE first identifies which functions, types, variables and macros are covered by each test unit in a test suite. Each time the system under test is modified, TESTTUBE identifies which entities were change …	0.96
	Many selective regression testing strategies , which select a subset of the original tests for regression testing, have been proposed recently. Yet, the basic assumptions supporting selective regression testing have not been fully examined. The paper first introduces the notion of scope for change to represent the regression testing focus. A classification …	0.97
	Regression testing is a necessary but expensive activity aimed at showing that code has not been adversely affected by changes. A selective approach to regression testing attempts to reuse tests from an existing test …	0.97

Fig. 3. Similarity Results. We calculate the similarity between paper title and abstract.

would be have a high correlation in the embedding space.

H2: Each paper p can be represented as linear combination of certain latent base topics $t^i; i \in [1,n]$, where t^i's are learned using a topic-modelling approach like latent dirichlet allocation (LDA) [2]. Intuitively, one can imagine the information within p to be a pie, where topic t_p^i represents a slice of the pie.

H3: Given a query paper q, a set of candidate publications $\mathbf{P} = \{p_1, p_2, \ldots, p_k\}$ are considered "relevant" if: i) for each i, $\mathbf{e}_{p_i}$ and $\mathbf{e}_q$ are similar in the high dimensional embedding space, or ii) for each topic $t^i \in q$, there is atleast one paper $p_i \in \mathbf{P}$ that contains topic t^i, i.e. the set of relevant papers $\mathbf{P}$ span the space of topics in q.

The details on how these hypotheses translate to the final approach are described in the later sections.

3 Dataset

There is a large selection of publications freely available online. Even though one could technically scrape this data from the web, for the sake of feasibility, we instead use the publicly available DBLP Citation Network dataset [11, 12]. In its 10^{th} version, DBLP Citation Network has about $3,079,007$ papers and $25,166,994$ citation relationships. For each paper, it contains the following fields: paper id, paper title, authors, venue, year, number of citations, references, and abstract. Figure 2 highlights a couple of examples from the dataset. The references can be used to form a directed network of papers within the dataset. Another advantage of using this dataset is the minimal amount of data wrangling required as the dataset is already curated .

As shown in Figure 1, we want our proposed tool to seamlessly transition between the different use-cases. We now describe our initial thoughts on the possible design choices, and how the data abstractions defined in Section 4 are used in the visualization. An overview of our proposed InfoVis solution is shown in Figure 4.

4 Data and Task Abstractions

As shown in Figure 1, our proposed approach has three use cases: i) providing an overview of all the paper, ii) recommending relevant publications, and iii) allowing trend analysis via a chronological view. To realise these effectively, we need to abstract the data to obtain relevant characteristics. We now discuss these abstractions.

4.1 High Dimensional Embedding

As mentioned in Section 2, **H1** assumes that each paper p can be represented as a high dimensional embedding $\mathbf{e}_p$. We envision using the title and abstract fields of each paper p to obtain $\mathbf{e}_p$. An off-the-shelf pre-trained language model [1] could be used to embed the title/abstract in to a high dimensional semantic space. As the pre-trained language model is trained of a generic web dataset, there is a possibility that the embeddings it generates aren't specific to our task involving research papers. Therefore, training a language model on the DBLP Citation Network dataset would be an avenue for exploration.

4.2 Topic Modeling

Hypothesis **H2** presumes that each paper p can be represented using several topics t^i. The goal here is to identify a collection of words,

[1] Such as `https://spacy.io/models/en-starters`

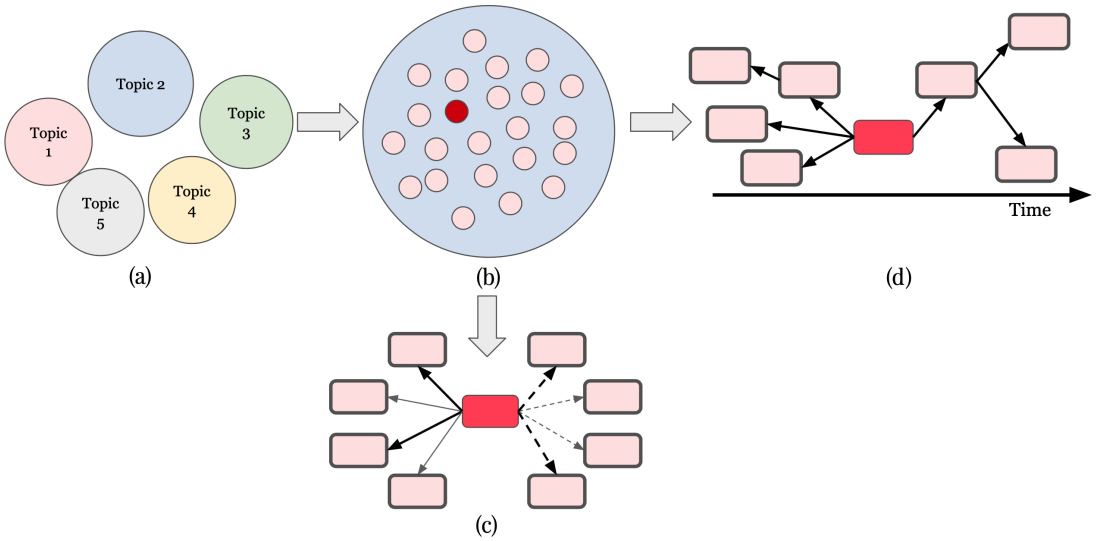

Fig. 4. **Proposed InfoVis solution**. (a) Grouped topics, (b) Zoomed-in View, (c) Recommend set of papers, (d) Chronological view for trend analysis.

which grouped together, constitute a particular topic in p. Using the title/abstract for all papers in the dataset, we aim to leverage a probabilistic topic modeling approach called Latent Dirichlet Allocation (LDA) [2] to obtain the set of topics $t^i; i \in [1,n]$. Applying this trained model, for each paper p, we can obtain a list of scores $\alpha_p = [\alpha_1^i, \ldots, \alpha_n^i]$ where $\alpha_p^i \in [0,1]$ indicates the probability of topic t^i being present in p.

4.3 Relevance Algorithm

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, **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,

$$s_{pq} = \frac{\mathbf{e}_p \odot \mathbf{e}_q}{||\mathbf{e}_p|| \cdot ||\mathbf{e}_q||} \tag{1}$$

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 $\alpha_q = [\alpha_1^q, \ldots, \alpha_n^q]$ be the topic scores computed for a query paper q. For each $\alpha_i^q > \tau$, where τ is a positive threshold, our aim then is to find a publication p_i that: i) has a high $\alpha_i^{p_i}$ value indicating that topic t^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.

5 Proposed Implementation Solutions

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 [6]. 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 4(a). Following hypothesis **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 4(b). Each small circle in Figure 4(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 4(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 4(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 4(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.

6 Related Work

From visualization perspective, a 2D scatter plot of papers formed using a dimensionality reduction technique such as t-SNE [6] or UMAP [7] is not new and has been used extensively to visualize related papers where closeness in 2D space represents similarity [1,5]. 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 [5] 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* [4], where they proposed a tree-based visualization of hierarchically clustered documents intended for search and exploration.

A related CPSC 547 course project *PaperQuest* [10] 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 [9] or Word2Vec [8] 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 [3] 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.

- Raghav: High dimensional embedding and InfoVis solution.
- Shih-Han: High dimensional embedding and InfoVis solution.
- Siddhesh: Topic modeling and InfoVis solution.

We aim to be done with all the data processing and abstraction by November 10^{th}, after which we will focus on the implementing the visualization.

REFERENCES

[1] Iclr paper explorer. `https://iclr.cc/virtual_2020/paper_vis.html`. Accessed: 2020-10-22.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[3] M. Bostock, V. Ogievetsky, and J. Heer. D^3 data-driven documents. *IEEE transactions on visualization and computer graphics*, 17(12):2301–2309, 2011.

[4] M. Brehmer, S. Ingram, J. Stray, and T. Munzner. Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists. *IEEE transactions on visualization and computer graphics*, 20(12):2271–2280, 2014.

[5] A. Crisan, T. Munzner, and J. L. Gardy. Adjutant: an r-based tool to support topic discovery for systematic and literature reviews. *Bioinformatics*, 35(6):1070–1072, 2019.

[6] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 2008.

[7] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[8] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 2013.

[9] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[10] A. Ponsard and F. Escalona. Paperquest: a visualization tool to support literature review. `https://www.cs.ubc.ca/~tmm/courses/547-14/projects/antoine-pax/report.pdf`, 2014.

[11] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. Hsu, and K. Wang. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, 2015.

[12] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008.