Knowledge Discovery in Text Corpora with Diffusion Maps and Information Theory

Oliver Kirsten

July 24, 2025

1 Introduction

The analysis of document complexity and semantic relationships in large text corpora requires methods that can capture both the hierarchical nature of knowledge and the asymmetric dependencies between concepts. Traditional text mining approaches treat document similarity as a symmetric problem, failing to model the fundamental asymmetries inherent in knowledge representation and learning pathways.

This work introduces a novel framework that combines information-theoretic measures with diffusion geometry to reveal the intrinsic complexity structure of document collections. Our approach leverages cross-entropy kernels to capture asymmetric semantic relationships and SVD-based entropy measures to characterize document complexity along a generality-specialization spectrum.

1.1 Problem Statement

The central challenge in knowledge discovery from text corpora lies in understanding how documents relate to each other not just in terms of topical similarity, but in terms of conceptual complexity and prerequisite relationships. Traditional similarity measures face several critical limitations: symmetric distance metrics cannot capture the directional nature of knowledge dependencies (e.g., that advanced concepts require foundational knowledge but not vice versa), existing complexity measures rely on surface-level features like document length rather than semantic depth, standard clustering approaches fail to preserve the hierarchical structure inherent in knowledge domains, and current methods lack principled approaches for discovering learning pathways between conceptually distant topics.

These limitations are particularly pronounced in educational and technical domains where understanding complexity gradients and conceptual dependencies is crucial for effective knowledge organization.

1.2 Proposed Approach

Our framework addresses these challenges through three key ideas. First, we introduce SVD entropy as a compression-based measure of document complexity that distinguishes between general foundational content and specialized technical material based on vocabulary distribution patterns in the semantic space. Second, we develop asymmetric

1

cross-entropy kernels that measure the information cost of encoding one document using another's vocabulary distribution, naturally capturing directional semantic relationships and prerequisite dependencies. Third, we combine these information-theoretic measures with diffusion geometry to compute semantic geodesics that represent optimal learning pathways through the knowledge space, respecting both complexity gradients and conceptual relationships.

This approach enables the discovery of natural progression sequences in knowledge domains while providing quantitative characterization of document complexity.

2 Data

2.1 Dataset Description

The dataset used in this study is derived from the OpenWeb-Math [5] corpus, specifically filtered and curated for mathematical machine learning content. The data collection and preprocessing pipeline involved several stages:

Source Selection: Documents were extracted from the OpenWeb-Math dataset with URL filtering restricted to sources including "wikipedia.org", "github.io", and "nature.com".

Topic-based Retrieval: Zero-shot BERTopic [6] was employed to retrieve documents matching specific mathematical and machine learning topics, including linear algebra, statistics, probability theory, information theory, machine learning, deep learning, statistical learning, reinforcement learning, and neural networks.

Preprocessing Pipeline: Retrieved documents underwent standard NLP preprocessing techniques including tokenization, stopword removal, and stemming to prepare the text for subsequent analysis.

Each document contains text content, URL, and topic classification, with topics ranging from basic concepts like "Introducing matrices" to advanced topics like "LSTM networks".

2.2 Data Properties

This dataset provides an ideal testbed for evaluating semantic relationships and learning pathways in mathematical and machine learning concepts. The hierarchical nature of mathematical knowledge, from foundational linear algebra to advanced neural architectures, allows us to validate whether our diffusion geodesics capture meaningful semantic progressions.

The documents vary significantly in complexity and abstraction level, providing a natural gradient for testing our entropy-based characterization methods. Furthermore, the well-defined conceptual relationships in mathematics and machine learning enable qualitative validation of discovered semantic paths.

3 Methods

3.1 Text Preprocessing and Vectorization

Our pipeline begins with text preprocessing and Frequency-Inverse Document Frequency (TF-IDF) [2] vectorization:

$$\mathbf{X}_{t,d} = \text{TF}(t, d) \cdot \log \left(\frac{N}{|\{d' : t \in d'\}|} \right) \tag{1}$$

where t represents a term, d a document, and N the total number of documents. The preprocessing includes removal of terms appearing in more than 50% of documents and terms appearing in fewer than 5 documents, along with Gensim preprocessing for text cleaning.

3.2   Singular Value Decomposition

To manage computational complexity while preserving semantic information, we apply SVD:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{2}$$

where $\mathbf{X} \in \mathbb{R}^{n \times m}$ is the TF-IDF matrix, and we retain the top $k = 5000$ components. This decomposition is also known as Latent Semantic Analysis [3], and provides dimensionality reduction from vocabulary size to manageable dimensions, noise reduction through low-rank approximation, and computational efficiency for subsequent operations.

3.3   Non-negative SVD (NN-SVD) Transformation

Since diffusion processes require non-negative values, we apply a Non-negative SVD (NN-SVD) [4] transformation that preserves the dominant singular vector components while ensuring non-negativity. This transformation also serves as an effective initialization strategy for Non-negative Matrix Factorization (NMF).

The NN-SVD approach addresses the fundamental challenge that standard SVD produces factors with both positive and negative entries, while many applications (including diffusion processes and NMF) require non-negative representations. For each singular vector component, we preserve the orientation that captures the majority of the signal energy:

$$\mathbf{U}_{\text{nonneg}}[:, i] = \begin{cases} \max(\mathbf{U}[:, i], \epsilon) & \text{if } \|\mathbf{U}_+[:, i]\|_2 \geq \|\mathbf{U}_-[:, i]\|_2 \\ \max(-\mathbf{U}[:, i], \epsilon) & \text{otherwise} \end{cases} \tag{3}$$

where $\mathbf{U}_+$ and $\mathbf{U}_-$ denote positive and negative components respectively, and $\epsilon = 10^{-5}$ prevents numerical issues.

This transformation strategy offers several advantages. Energy preservation is achieved by selecting the orientation with larger L_2 norm, retaining the maximum amount of information from each singular vector while satisfying non-negativity constraints. The approach maintains the semantic structure captured by SVD while enabling probabilistic interpretations required for diffusion analysis.

3.4   Entropy-based Similarity Kernel

The NN-SVD matrix fundamentally serves as a change of basis for the original TF-IDF matrix, projecting the document representations from the high-dimensional vocabulary space into a lower-dimensional semantic space with non-negative coordinates.

The NN-SVD provides a new basis where each document is represented as:

$$\mathbf{X}_{\text{nonneg}} = \mathbf{U}_{\text{nonneg}}^T \mathbf{X} \tag{4}$$

where $\mathbf{U}_{\text{nonneg}}$ forms the new basis vectors derived from the non-negative transformation of the left singular vectors.

We normalize the transformed non-negative matrix to form probability distributions:

$$\mathbf{P} = \frac{\mathbf{X}_{\text{nonneg}}}{\sum_i \mathbf{X}_{\text{nonneg}}[i,:]} \tag{5}$$

and compute a cross-entropy based kernel:

$$\mathbf{K}_{ij} = -\sum_k \mathbf{P}_{ik} \log_2(\mathbf{P}_{jk}) \tag{6}$$

3.4.1 SVD Entropy as Document Generality and Complexity Measure

The SVD entropy provides a novel characterization of document generality through the fundamental principle that **documents using common, widely-shared vocabulary can be more efficiently compressed using the dominant SVD components, while documents with specialized vocabulary require more components for accurate representation**.

$$H_{\text{SVD}}(d) = \frac{\mathbf{K}_{dd}}{\log_2(n)} \tag{7}$$

This measure captures document characteristics along multiple dimensions. Documents with low SVD entropy exhibit high generality, extensively using vocabulary patterns captured by the top SVD components. These documents employ common terminologies and conceptual frameworks found across the corpus, typically representing introductory material, overviews, or foundational concepts that can be efficiently "explained" using the dominant semantic patterns in the corpus. Examples include basic tutorials, general introductions, and widely-applicable principles.

Conversely, documents with high SVD entropy demonstrate high specialization, requiring many SVD components for accurate reconstruction. These documents use terminology and conceptual patterns that are rare or unique within the corpus, representing advanced, specialized, or highly technical content that cannot be efficiently compressed using only the dominant corpus-wide patterns. Examples include research papers, technical specifications, and domain-specific advanced discussions.

The compression-based intuition reveals that SVD entropy directly measures how "compressible" a document is using the corpus-wide semantic patterns. Just as simple images with common features compress well with few JPEG coefficients while complex images with unique details require many coefficients for faithful reproduction, general documents using common vocabulary patterns compress well with few SVD components (low entropy) while specialized documents with unique terminology require many SVD components (high entropy).

This measure has important pedagogical implications, automatically identifying documents by their position in the learning hierarchy. Low entropy documents often represent accessible entry points, while high entropy documents represent advanced material requiring substantial background knowledge. The SVD entropy thus serves as an automatic

"difficulty level" indicator based on vocabulary specialization rather than subjective assessments.

This compression-based complexity measure provides an objective, corpus-relative assessment of document generality that scales naturally with the semantic diversity of the collection.

3.4.2 Asymmetric Cross-Entropy Kernel Interpretation

The cross-entropy kernel $\mathbf{K}_{ij}$ has a profound information-theoretic interpretation that captures the fundamental asymmetries in semantic relationships between documents. Specifically, $\mathbf{K}_{ij}$ measures the **expected number of bits required to encode document i using the vocabulary distribution of document j as the coding scheme**.

This asymmetric formulation reveals several important semantic phenomena. Information flow and teaching relationships emerge when $\mathbf{K}_{ij} < \mathbf{K}_{ji}$, indicating that document j provides a more efficient encoding for document i than vice versa. This suggests that j contains the "vocabulary foundation" or conceptual framework needed to understand i, and in educational contexts, this indicates that j could serve as a prerequisite or explanatory text for i.

Generalization versus specialization asymmetries become apparent when document j is a comprehensive overview and document i is specialized. In such cases, $\mathbf{K}_{ij}$ (encoding specialist using generalist's vocabulary) will typically be lower than $\mathbf{K}_{ji}$ (encoding generalist using specialist's vocabulary). This reflects the natural relationship where general texts provide broad vocabulary that can describe specialized concepts, but specialized texts lack the breadth to efficiently describe general concepts.

The asymmetry also enables conceptual dependency detection, capturing which documents can "explain" others through shared conceptual vocabulary. Documents with lower cross-entropy when used as encoders likely represent foundational or prerequisite knowledge. Additionally, documents that serve as bridges between domains will exhibit asymmetric relationships with texts from both domains, potentially identifying interdisciplinary or survey documents.

This information-theoretic perspective transforms document similarity from a symmetric geometric problem into an asymmetric knowledge flow analysis, revealing the directional relationships inherent in learning pathways and conceptual dependencies.

3.5 Diffusion Process Construction

We construct a similarity matrix using a Gaussian kernel applied to the entropy-based divergences:

$$\mathbf{S}_{ij} = \exp\left(-\frac{\mathbf{K}_{ij}}{\sigma^2}\right) \tag{8}$$

where $\sigma = 1.5$ is a bandwidth parameter. The similarity matrix is then row-normalized:

$$\mathbf{M}_{ij} = \frac{\mathbf{S}_{ij}}{\sum_k \mathbf{S}_{ik}} \tag{9}$$

forming a column-stochastic Markov transition matrix.

3.6 Eigendecomposition and Diffusion Distance

We compute the eigendecomposition of the transpose of the Markov matrix:

$$\mathbf{M}^T \mathbf{v}_i = \lambda_i \mathbf{v}_i \tag{10}$$

With this eigendecomposition, the simulated diffusion (random walks) of t steps on $\mathbf{M}$ can be efficiently computed:

$$\mathbf{M}^t = \mathbf{V} \text{diag}(\lambda_1^t, \ldots, \lambda_k^t) \mathbf{V}^{-1} \tag{11}$$

For non-integer values of t and negative eigenvalues, the complex valued root is calculated and only the real part of that is kept.

The diffusion distance between points i and j at time t is defined as:

$$D_t(i, j) = \sqrt{\sum_k \frac{(\mathbf{M}^t[i, k] - \mathbf{M}^t[j, k])^2}{\pi_k}} \tag{12}$$

where π_k is the stationary distribution of the Markov matrix $\mathbf{M}$.

3.7 Geodesic Path Finding

We construct a k-nearest neighbors graph using diffusion distances and apply Dijkstra's algorithm [7] to find the shortest paths:

Algorithm 1 Geodesic Path Finding

   Input: Diffusion distances $\mathbf{D}_t$, parameter k
   Output: Shortest path between source and target
   for each document i **do**
      Find k nearest neighbors based on $\mathbf{D}_t[i, :]$
      Add edges to KNN graph with diffusion distances as weights
   end for
   Apply Dijkstra's algorithm for shortest path computation

---

4 Results

4.1 Geodesic Path Analysis

We computed geodesic paths between conceptually distant documents, starting from "Introducing matrices" (basic linear algebra) and ending at "LSTM" (advanced neural networks). The analysis reveals several key insights about semantic structure and learning pathways in the mathematical machine learning domain.
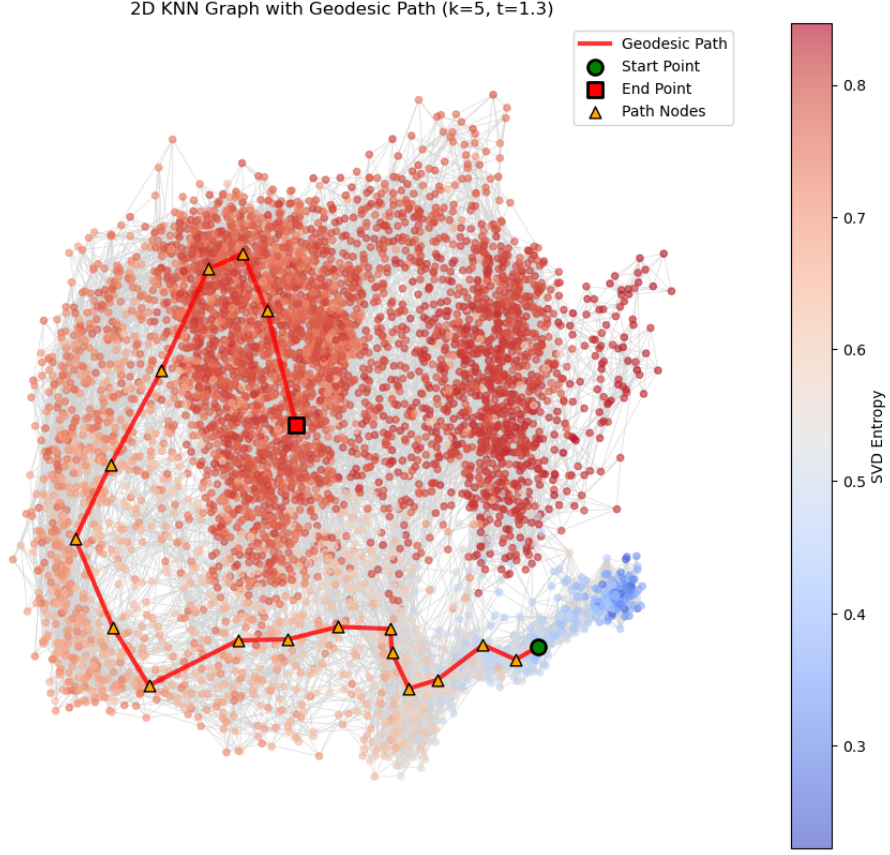
4.2 Visualization Analysis

Figure 1: Geodesic path visualization showing semantic transitions between documents in the diffusion map embedding.

The 2D spring layout [8] visualization in Figure 1 reveals the intrinsic geometric structure of the document manifold through diffusion map embedding. The color gradient represents SVD entropy values, creating a natural "complexity landscape" where blue regions (low entropy, 0.3-0.4) may correspond to foundational concepts, and red regions (high entropy, 0.7-0.8) specialized topics.

The geodesic path (red line with triangular markers) demonstrates a clear semantic progression from the bottom-right (start point, green circle) to the center-left (end point, red square). This trajectory traverses multiple entropy gradients, suggesting the path naturally follows increasing conceptual complexity. Notably, the path avoids the densest red regions, indicating that the algorithm discovers routes through intermediate complexity levels rather than attempting direct transitions between fundamentally different concept spaces.

4.2.1 Geodesic Path Progression Analysis

Table 1 shows the semantic progression along the discovered 19-step path, exhibiting a clear entropy gradient from 0.4861 (basic matrices) to 0.7738 (LSTM networks).

The path reveals a continuous progression through the conceptual landscape, beginning with basic mathematical foundations and gradually transitioning through core ma-

Table 1: Geodesic Path from Start to End Point

Index	Text (first 125 chars)	SVD Entropy
0	Introducing matrices Published: Here, I will introduce the three main ways of thinking about matrices. This high-level de ...	0.486100
1	Data for Exercise 7.40 Backtoback Format A data frame/tibble with 24 observations on two variables score a numeric ve ...	0.566400
2	Autoencoder Difference between generative and discriminative modelling Generative modelling Generative modelling ...	0.523300
3	Measure the uncertainty in deep learning models using dropout Seminal blog post of Yarin Gal from Cambridge machine learni ...	0.626900
4	In today's post I would like to give you a quick-and-dirty introduction into a neural network architecture type called Autoen ...	0.569500
5	Conv Layers In turn, each neuron in the second conv layer is connected only to neurons located within a small rectangle in t ...	0.596300
6	Chapter 2: Matrices and Linear Algebra¶ import igl import scipy as sp import numpy as np from meshplot import plot, subplo ...	0.584800
7	A practical introduction to GNNs - Part 1 This is Part 1 of an introductory lecture on graph neural networks that I gave f ...	0.618200
8	Generative Adversarial Network (GAN) in TensorFlow - Part 4 The GAN Class and Data Functions Now that we're able to im ...	0.618800
9	Lecture 6: Further Examples of Classifiers¶ This week we're discussing more classifiers and their applications. Suppor ...	0.699200
10	K-Nearest Neighbor from Scratch in Python Posted by Kenzo Takahashi on Wed 06 January 2016 We are going to implement K-ne ...	0.688400
11	Multi-layer Perceptron (MLP)¶ New in version 0.3. In this module is stored everything related to Multi-layer perceptron (...	0.641500
12	TensorFlow 101 Word2Vec Shan-Hung Wu DataLab Fall 2018 TensorFlow is a powerful open source libraray used for large-s ...	0.722100
13	Libraries import numpy as np import pandas as pd import matplotlib.pyplot as plt import seaborn as sns sns.set() matplo ...	0.678100
14	2.1 Three Popular Data Displays Learning Objective 1. To learn to interpret the meaning of three graphical represent ...	0.728600
15	Machine learning For the journal, see Machine Learning (journal). Machine learning is a subfield of computer science and ...	0.723500
16	Exploring and Understanding Hyperparameter Tuning Learners use hyperparameters to achieve better performance on particular ...	0.765100
17	In this post I'll talk through the pieces of the tensorflow API most relevant to building recurrent neural networks. The tens ...	0.771900
18	LSTMs 2 minute read LSTMs to Model Physiological Time Series Harini Suresh, Nicholas Locascio, MIT Neural Net ...	0.773800

chine learning concepts including autoencoders and neural networks, ultimately reaching sophisticated architectures like LSTM networks.

The smooth entropy progression with minimal backtracking demonstrates that diffusion geodesics successfully identify natural learning sequences, automatically discovering pedagogically meaningful pathways.

5 Discussion

5.1 Methodological Advantages

Our diffusion-based methodology offers several significant advantages over traditional text mining approaches. By treating the text corpus as a manifold with intrinsic geometry, we enable natural distance computations that reflect semantic relationships. The diffusion process naturally incorporates a time parameter, allowing analysis at multiple scales and the discovery of hierarchical relationships. The use of entropy-based similarity measures and diffusion distances ensures that the intrinsic semantic relationships are preserved throughout the analysis. Optimizations using Numba and efficient data structures enable the processing of medium-scale document collections with reasonable computational resources. The geodesic paths and diffusion map embeddings provide results that are not only effective for analysis but also interpretable in terms of semantic relationships and document complexity.

5.2 Limitations and Future Work

While our methodology demonstrates promising results, several limitations warrant consideration. The current implementation is suitable for medium-scale corpora, but scaling to very large document collections would require distributed computing approaches or approximate methods for eigendecomposition and distance computations. The method involves several parameters (SVD components, diffusion time, kernel bandwidth) that may require tuning for different domains and corpus characteristics. Additionally, the approach has been validated primarily on mathematical and machine learning content, and evaluation on other domains would strengthen the generalizability claims.

Future work could explore adaptive parameter selection, distributed implementations for large-scale corpora, and validation across diverse text domains to establish broader applicability.

6 Conclusion

This work presents a novel framework for text mining that leverages diffusion geometry and information theory to reveal the underlying semantic structure of document corpora. Our approach successfully combines TF-IDF vectorization, SVD-based dimensionality reduction, entropy-based similarity measures, and diffusion processes to create a comprehensive analysis pipeline.

The key innovations include the development of SVD entropy as a measure of document generality and complexity, the use of asymmetric cross-entropy kernels to capture directional semantic relationships, and the computation of semantic geodesics that represent meaningful learning pathways between concepts.

Evaluation on a mathematical machine learning corpus demonstrates the method's effectiveness in discovering semantic relationships, characterizing document complexity, and finding interpretable paths between conceptually distant topics. The approach provides both quantitative measures through diffusion distances and qualitative insights through geodesic path analysis.

The framework opens new possibilities for educational technology, where semantic geodesics could guide personalized learning paths, and for knowledge discovery, where the geometric structure of document collections can reveal hidden relationships and progression patterns in specialized domains.

References

[1] Coifman, R. R., & Lafon, S. (2006). Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1), 5-30.

[2] Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513-523.

[3] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391-407.

[4] Boutsidis, C., & Gallopoulos, E. (2008). SVD based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4), 1350-1362.

[5] Paster, K., et al. (2023). OpenWebMath: An open dataset of high-quality mathematical web text. *arXiv preprint arXiv:2310.06786*.

[6] Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

[7] Dijkstra, E. W. (2022). A note on two problems in connexion with graphs. In *Edsger Wybe Dijkstra: His Life, Work, and Legacy* (pp. 287-290).

[8] Kobourov, S. G. (2012). Spring embedders and force directed graph drawing algorithms. *arXiv preprint arXiv:1201.3011*.