Scalable Alternatives to Singular Value Decomposition for Latent Semantic Indexing

Vishva Bhavsar, Kandarp Malkan, Srinibas Masanta and Jaspreet Kaur Ghotra

Abstract

Singular Value Decomposition (SVD) is a widely used technique for Latent Semantic Indexing (LSI). However, it becomes computationally expensive for large datasets. This study explores alternative low-rank approximation algorithms, such as Randomized SVD (RSVD) and Lanczos SVD, to efficiently compute LSI on large-scale datasets. Using the CNN/Daily Mail dataset, we evaluate these methods based on runtime and their scalability for handling large data. The results highlight the trade-offs between computational complexity and scalability, providing insights into efficient solutions for information retrieval tasks.

Keywords

Latent Semantic Indexing, Singular Value Decomposition, Randomized SVD, Lanczos SVD, Information Retrieval

1. Problem

Singular Value Decomposition (SVD) serves as the backbone for Latent Semantic Indexing (LSI) in information retrieval. Despite its effectiveness, SVD becomes computationally prohibitive in terms of memory and time when dealing with large datasets. This limitation restricts its applicability in real-world scenarios where large-scale text corpora are common. The challenge lies in identifying alternative low-rank approximation methods that can reduce computational complexity and memory requirements while preserving the semantic quality of LSI. This study focuses on two promising methods: Randomized SVD (RSVD) and Lanczos SVD.

2. Dataset

The study uses the CNN/Daily Mail dataset's train split, comprising 287,113 articles. Key attributes include:

- **Articles**: Full text of the news articles.
- **Highlights**: Summarized content.
- **ID**: Unique identifier for each article.

The TF-IDF matrix is constructed with dimensions (287,113 documents × 639,030 unique terms).

Introduction to Information Retrieval (IT550), Autumn 2024, DAIICT, India

202318019@daiict.ac.in (V. Bhavsar); 202318021@daiict.ac.in (K. Malkan); 202318054@daiict.ac.in (S. Masanta); 202318058@daiict.ac.in (J. K. Ghotra)

202318019 (V. Bhavsar); 202318021 (K. Malkan); 202318054 (S. Masanta); 202318058 (J. K. Ghotra)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

3. Evaluation Metrics

The primary evaluation metric for this study is runtime, reflecting the computational efficiency of the algorithms. Given the high dimensionality of the dataset, runtime serves as a critical measure of scalability.

4. Results

The two alternative algorithms, Randomized SVD (RSVD) and Lanczos SVD, were applied to the TF-IDF matrix to compute the top 100 singular values and vectors. Their runtime performances and output shapes are summarized below:

4.1. Randomized SVD

- **Runtime:** 145.93 seconds
- **Output Shapes:**
  - U (Document-Concept Similarity Matrix): (287,113, 100)
  - S (Concept Strengths): (100,)
  - V^T (Term-Concept Similarity Matrix): (100, 639,030)

4.2. Lanczos SVD

- **Runtime:** 226.34 seconds
- **Output Shapes:**
  - U (Document-Concept Similarity Matrix): (287,113, 100)
  - S (Concept Strengths): (100,)
  - V^T (Term-Concept Similarity Matrix): (100, 639,030)

5. Key Challenges and Learnings

5.1. Key Challenges

- Identifying alternative low-approximation algorithms to replace SVD required extensive research, and understanding the mathematical foundations of Randomized SVD and Lanczos algorithms added complexity.
- Adapting these algorithms for text data in LSI involved significant trial and error, while managing and preprocessing the large dataset posed resource-intensive challenges.

5.2. Learnings

- Learned how Randomized SVD and Lanczos methods provide efficient approximations for large-scale matrix decomposition, developed a mathematical understanding of their workings, and gained expertise in applying these concepts to text data for dimensionality reduction in LSI.
- Discovered how to implement these algorithms for large datasets, making LSI practical for large-scale applications.

{202318019, 202318021, 202318054, 202318058}@daiict.ac.in
©2024. Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0).

6. Conclusion

This study highlights the potential of alternative low-rank approximation techniques, specifically RSVD and Lanczos SVD, in addressing the scalability challenges of LSI for large-scale datasets. While both methods demonstrated the ability to compute latent structures effectively, RSVD emerged as the more efficient option in terms of runtime. These findings underscore the importance of exploring scalable algorithms for real-world information retrieval applications.

7. References

- Brunton, Steven L., and J. Nathan Kutz. Data Driven Science & Engineering: Machine Learning, Dynamical Systems, and Control. Cambridge: Cambridge University Press, 2019.
- Olney, A. M. Large-scale latent semantic analysis.

{202318019, 202318021, 202318054, 202318058}@daiict.ac.in
©2024. Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0).