Latent Semantic Indexing: A Probabilistic Analysis

Seminar “Theoretical Topics in Data Science”

Vahe Eminyan

October 23, 2023
RWTH Aachen University

Latent semantic indexing (LSI) is one of the techniques used for information retrieval. LSI uses a mathematical technique called Singular Value Decomposition (SVD) applied to the term-document matrix. It is used to discover the hidden (latent) structure in the text data.

LSI is widely used in practice and has shown strong empirical results. However, due to the large size of term-document matrices, the computation can be slow. To expedite computations, the paper suggests initially mapping the large matrices into low-dimensional spaces using random projections and then applying LSI to this reduced-dimensional projection. The paper proves that the matrix obtained via random projection followed by LSI recovers almost as much as the matrix obtained by direct LSI, with high probability.

1 Introduction

In this section, we will introduce the topic and explain how the paper is structured. My plan is to cover the sections 1,2,5,6 of the paper.

- In sec2: What is LSI and why to use it
- In sec3: The computation with large matrices takes too long. We need random projections to speed up the process.
- In sec3: prove that the matrix obtained via random projection followed by LSI recovers almost as much as the matrix obtained by direct LSI, with high probability.
- In sec3: Conclusion and short summary

2 LSI Background

It is important to understand the meaning of LSI, how it works, and its mathematical background (SVD).

3 LSI by Random Projection

In order to reduce the computational time we use dimensionality reduction and then apply LSI. We prove theorem 5 of the paper which maintains that the matrix obtained via random projection followed by LSI recovers almost as much as the matrix obtained by direct LSI, with high probability.

4 Conclusion and Summary

In this section, we will draw conclusions based on the findings presented in the seminar paper and give a brief summary of the paper.

References