Latent Semantic Indexing: A Probabilistic Analysis

Seminar "Theoretical Topics in Data Science"

Vahe Eminyan

October 24, 2023
RWTH Aachen University

This seminar paper gives an analysis of the scientific paper "Latent Semantic Indexing: A Probabilistic Analysis" by Papadimitriou et al.
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. These lead to two interesting questions: How can we speed up computations, and why does LSI perform so well?
Papadimitriou et al. address both of these questions. To expedite computations, they suggest initially mapping the large matrices into low-dimensional spaces using random projections and then applying LSI to this reduced-dimensional projection. They 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. Additionally, they provide a theoretical justification for LSI's effectiveness, considering a special type of term-document matrix.
In this seminar paper, we mainly focus on the first question, providing a detailed analysis of LSI by random projection.

1 Introduction

In this section, we will introduce the topic and explain how the paper is structured. My plan is to cover all sections focusing on section 5.

2 LSI Background

In this part we will explain the technique LSI, how it works, and its mathematical background (SVD).

3 Probabilistic Corpus Model

In this section, we will introduce the Probabilistic Corpus Model.

4 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.

5 Brief Mention of Analysis of LSI's Good Performance

In this part, I will provide the most significant information of paragraph 4 from the original paper. Main idea: Under certain conditions and assumptions, it can be shown why LSI performes well.

6 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