Latent Semantic Indexing

Seminar “Theoretical Topics in Data Science”

Vahe Eminyan

vahe.eminyan@rwth-aachen.de

18.11.2023

Overview

Motivation

- Large datasets, often organized in tabular form, represented as matrices
  - Term-document matrix representing word occurrence in documents
  - Movie-user matrix representing watched movies of users
- Interesting aspects
  - Find documents semantically associated with a query
  - Recommend a new movie to a user

	Doc 1	Doc 2	...	Doc m
Term 1	0	1	...	1
Term 2	1	0	...	1
...	...	...	...	...
Term n	1	0	...	0

Documents

Terms $\begin{pmatrix} 0 & 1 & \dots & 1 \\ 1 & 0 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \end{pmatrix}$

$n \times m$

Latent Semantic Indexing

- LSI as an information retrieval method
- Finds the latent (hidden) semantic structure of textual data
- Represent term-document matrix as product of three matrices: term-topic, topic-topic and topic-document matrix
- Answer queries with help of these matrices
- Based on singular value decomposition of the matrix

In this section

- SVD explanation
- How does LSI work

Singular Value Decomposition (SVD)

Original Paper Overview and Emphasized Aspect

In this section

- Two interesting questions Papadimitriou et al. investigated [1]
  - Why does LSI perform well (why does it find the documents semantically related to each other)
  - How can we speed up the computation
- We will focus on the second question

In this section

- In this section we will investigate the question ”How we can speed up the computation”: Informal formulation of the main theorem of this section (Theorem 5 original paper)
- Introduction of theorems and lemmas that are necessary for the proof of the main theorem
- Introduction: the main theorem (Theorem 5 original paper)
- Proof of the main theorem (Theorem 5 original paper)
- Computational savings achieved by LSI by random projection

References

Christos H. Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, and Santosh Vempala.
Latent semantic indexing: A probabilistic analysis.
Journal of Computer and System Sciences, 61(2):217–235, 2000.
URL: https://www.sciencedirect.com/science/article/pii/S0022000000917112, doi:10.1006/jcss.2000.1711.

References

The End