

Latent Semantic Indexing: A Probabilistic Analysis

Seminar "Theoretical Topics in Data Science"

Vahe Eminyan

November 3, 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: why does LSI perform so well, and how can we speed up computations? Papadimitriou et al. address both of these questions. They provide a theoretical justification for LSI's effectiveness, considering a special type of term-document matrix.

Additionally, 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.

In this seminar paper, we will consider all relevant information, providing a detailed analysis of LSI by random projection.

1 Introduction

Retrieving information from given data has been an important aspect for a long time. Due to the development of information systems and the Internet, huge amounts of data are available. Handling these huge amounts of data is a great challenge, so many techniques have been developed to cope with the work process. Consider a scenario where we have a dataset comprising millions of documents structured as a term-document matrix and a user submits a query. Instead of just finding the documents that include the words of the query, the goal is to find documents that align with the query semantically. This semantic understanding ensures more accurate and relevant search results, even in extensive datasets. Latent Semantic Indexing (LSI) is one of the information retrieval techniques. It uses Singular Value Decomposition (SVD) to represent the term-document matrix as a product of three matrices. By representing the term-document matrix in such a way we want to find the underlying (latent, hidden) semantical topics (also called concepts) of the term-document matrix. With the help of such decomposition, we can map the documents and queries to a lower dimensional space and compare them not only syntactically but also semantically. (Section 2 provides more detailed information about LSI and SVD).

LSI has shown strong empirical results. However, prior to the paper of Papadimitriou et al., there was no satisfactory explanation for its success (Why does LSI find the documents corresponding to the semantics of the query with high accuracy). They prove a theorem that under certain constraints LSI will always find the correct topic of the given query with high probability. Section 4 of this seminar paper elaborates on this theorem.

Despite its effectiveness, the computational time of LSI is very long. Papadimitriou et al. introduce a two-step LSI, in which we first map the original term-document matrix into a lower dimensional space by using random projections and then apply LSI. 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. Section 5 provides a detailed analysis of LSI by random projection and presents the proof for this statement.

In the last section, we will draw a conclusion and give a short summary of this seminar paper.

2 LSI Background

The corpus is defined as a set of documents. Each document is a vector of length n from \mathbb{R}^n . Each position of the document describes a mathematical function in terms of i th term of the entire term space. The function can, for example, be the frequency of the i th term, or just 0-1 (1 if the term appears in the document, 0 otherwise).

Let $A \in \mathbb{R}^{n \times m}$ be a matrix of rank r . The Singular Value Decomposition (SVD) is a mathematical technique that represents the matrix as a product of three matrices.

$$A = UDV^T.$$

Let $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r$ be the eigenvalues of the matrix AA^T (called singular values of A), then $D = \text{diag}(\sigma_1, \dots, \sigma_r) \in \mathbb{R}^{r \times r}$. $U \in \mathbb{R}^{n \times r}$ is the matrix representing the eigenvectors of the matrix AA^T and $V \in \mathbb{R}^{m \times r}$ the eigenvectors of the matrix $A^T A$. The columns of U and V are orthonormal. It is shown that every matrix can be represented in such a form [ZITATTTTTT]

In LSI the matrix $A \in \mathbb{R}^{n \times m}$ represents a term-document matrix of n terms and m documents. Each row of the matrix corresponds to a term and each column to a document.

LSI keeps only k -largest singular values of the matrix considering only the first k columns of matrices U and V . I.e.

$$A_k = U_k D_k V_k^T$$

which is an approximation of the original matrix A . In order to find the most similar documents for the submitted query we map the query to the k -dimensional space by using the matrix U_k . Afterward, we compare the k -dimensional projection with k -dimensional representations of the documents in the corpus. For this comparison, we can use cosine similarity. Thus we only need to store the three matrices U_k , D_k , and V_k . At this point, we can see the saving of the storage. In order to save the matrix A entirely we need to memorize $n \cdot m$ entries. In contrast to that, we need only $n \cdot k + k \cdot k + m \cdot k$ entries, which is much smaller than $n \cdot m$, because k is much smaller than both n and m .

The approximation A_k of rank k is the matrix that minimizes the Frobenius norm of the of $A - A_k$. Formally:

Theorem 1 (ZITATTT). *Among all $n \times m$ matrices C of rank at most k , A_k is the one that minimizes $\|A - C\|_F^2 = \sum_{i,j} (A_{ij} - C_{ij})^2$.*

I.e. this approximation preserves all relative distances, now we only need to say why it finds queries and documents that are semantically related.

3 Probabilistic Corpus Model

Because LSI relies on uncovering the statistical characteristics of a corpus, it is essential to begin with a well-defined probabilistic model of the corpus. In this section, we introduce the probabilistic corpus model. In the rest of the paper, we assume that the corpus is generated from a corpus model.

Definition 2. The universe U is the set of all terms.

Definition 3. A topic is a probability distribution on U .

The logical interpretation of this definition is the following. Assume we have a set of documents where one partition is about computer science and the other is about nature. Hence there would be two hidden topics for the whole set of documents. Hence the one topic will probably include terms like computer, software and algorithm, and the second will include the terms forest, ocean and earth. Thus we can interpret a topic as a probability distribution on the entire set of the terms.

Another important aspect is the authorship style, it influences the frequencies of words and hence the documents.

Definition 4. A style is a $|U| \times |U|$ stochastic matrix (a matrix with nonnegative entries and row sums equal to 1), denoting the way whereby style modifies the frequency of terms.

After introducing the previous three definitions, now we can define the probabilistic corpus model.

Definition 5. A corpus model \mathcal{C} is a quadruple $\mathcal{C} = (U, \mathcal{T}, \mathcal{S}, D)$, where U is the universe of terms, \mathcal{T} is a set of topics, and \mathcal{S} is a set of styles, and D a probability distribution on $\hat{\mathcal{T}} \times \hat{\mathcal{S}} \times \mathbb{Z}^+$, where by $\hat{\mathcal{T}}$ we denote the set of all convex combinations of topics in \mathcal{T} , by $\hat{\mathcal{S}}$ the set of all convex combinations of styles in \mathcal{S} , and by \mathbb{Z}^+ the set of positive integers (the integers represent the lengths of documents).

I.e. a corpus generated from the introduced corpus model is a set of documents in which every document is generated as follows: we sample a triple from D . It includes fixed $\hat{T} \in \hat{\mathcal{T}}$, $\hat{S} \in \hat{\mathcal{S}}$, and a fixed length ℓ of the document. Then we sample the terms of the document ℓ times according to \hat{T} .

4 Brief Mention of Analysis of LSI's Good Performance

In this section, we provide theorem that states under certain conditions the LSI brings similar documents together. First, we need a set of definitions for a corpus model.

Definition 6. A corpus model \mathcal{C} is pure if each document involves a single topic.

Definition 7. A corpus model \mathcal{C} is ε -seperable ($\varepsilon \in [0, 1)$), if a set of terms U_T is associated with each topic $T \in \mathcal{T}$ in the following way:

- U_T are mutually disjoint,
- for each T , the total probability T assign to the terms in U_T is at least $1 - \varepsilon$

We call U_T the primary set of terms of topic T .

In this section, we assume that the corpus model is style-free. Let $\mathcal{C} = (U, \mathcal{T}, D)$ be a pure and style-free corpus model and let $k = |\mathcal{T}|$ denote the number of topics in \mathcal{C} . Since \mathcal{C} is pure, each document generated from \mathcal{C} is in fact generated from some single topic T : hence we say that the document belongs to topic T . Let C be a corpus generated from \mathcal{C} and, for each document $d \in C$, let v_d denote the vector assigned to d by the rank- k LSI performed on C .

Definition 8. Rank- k LSI is δ -skewed on the corpus instance C if, for each pair of documents d and d' , $v_d \cdot v_{d'} \leq \delta \|v_d\| \|v_{d'}\|$, if d and d' belong to different topics and $v_d \cdot v_{d'} \geq 1 - \delta \|v_d\| \|v_{d'}\|$ if they belong to the same topic.

The following theorem shows that under some constraints on the corpus model the rank- k LSI indeed does well with high probability.

Theorem 9. Let \mathcal{C} be a pure, ε -seperable corpus model with k topics such that the probability each topic assigns to each term is at most τ , where $\tau > 0$ is a sufficiently small constant. Let C be a corpus of m documents generated from \mathcal{C} . Then, the rank- k LSI is $O(\varepsilon)$ -skewed on C with probability $1 - O(m^{-1})$.

In this seminar paper, we will omit the proof of this theorem. It can be found in the original paper.

5 LSI by Random Projection

In order to reduce the computational time we use dimensionality reduction and then apply LSI. According to Johnson and Lindenstrauss's lemma [ZiTAAAT] random projection of a matrix to a lower dimensional space preserves pairwise distances of its element while representing each element in lower dimensional space. However, it does not bring semantically connected elements (i.e. documents together). Papadimitriou et al. suggest a two-step LSI:

1. Apply a random projection to the initial corpus to ℓ dimensions, for some small $\ell > k$, to obtain, with high probability, a much smaller representation, which is still very close (in terms of distances and angles) to the original corpus.
2. Apply rank $O(k)$ LSI (because of the random projection, the number of singular values kept may have to be increased a little)

Before diving into the main theorem of this chapter, let's cover some formal groundwork. We again consider $A \in \mathbb{R}^{n \times m}$ which represents our term-document matrix and is generated from the corpus model. Let $R \in \mathbb{R}^{n \times \ell}$ be a random column-orthonormal matrix. We use R to project A into ℓ -dimensional space. We define $B := \sqrt{n/\ell} R^T A$ as the scaled random projection of A . The SVD representations of A and B are denoted as follows:

$$A = \sum_{i=1}^r \sigma_i u_i v_i \quad \text{and} \quad B = \sum_{i=1}^t \lambda_i a_i b_i.$$

We also introduce the following lemma with its corollary.

Lemma 10. Let ε be an arbitrary positive constant. If $\ell \geq c((\log n)/\varepsilon^2)$ for a sufficiently large constant c then, for $p = 1, \dots, t$

$$\lambda_p^2 \geq \frac{1}{k} \left[(1 - \varepsilon) \sum_{i=1}^k \sigma_i^2 - \sum_{j=1}^{p-1} \lambda_j^2 \right].$$

Corollary 11.

$$\sum_{p=1}^{2k} \lambda_p^2 \geq (1 - \varepsilon) \|A_k\|_F^2.$$

Theorem 12. Parsevals identity

Theorem 13. Eckart-Young-Minsky theorem.

Our rank- $2k$ approximation of matrix A is denoted as

$$B_{2k} := A \sum_{i=1}^{2k} b_i b_i^T$$

Now we can introduce the main theorem of this section.

Theorem 14.

$$\|A - B_{2k}\|_F^2 \leq \|A - A_k\|_F^2 + 2\varepsilon \|A\|_F^2$$

Informally, the theorem states that the original matrix A after applying random projection and then LSI is almost as good recovered as by using LSI on the original matrix.

Proof. We have

$$\begin{aligned} A &= \sum_{i=1}^r \sigma_i u_i v_i & A_k &= \sum_{i=1}^k \sigma_i u_i v_i \\ B &= \sum_{i=1}^t \lambda_i a_i b_i & B_{2k} &:= A \sum_{i=1}^{2k} b_i b_i^T. \end{aligned}$$

b_1, \dots, b_n is an orthonormal set of vectors spanning the row space of A , hence using the Parseval's identity we can write

$$\|A - B_{2k}\|_F^2 = \sum_{i=1}^n |(A - B_{2k})b_i|^2$$

for $i = 1, \dots, 2k$, because $b_i^T b_i = 1$ we have

$$(A - B_{2k})b_i = Ab_i - Ab_i = 0,$$

and for $i = 2k + 1, \dots, n$, because $b_j^T b_i = 0$ we have

$$(A - B_{2k})b_i = Ab_i.$$

Hence

$$\begin{aligned} \|A - B_{2k}\|_F^2 &= \sum_{i=1}^n |(A - B_{2k})b_i|^2 \\ &= \sum_{i=2k+1}^n |Ab_i|^2 = \sum_{i=1}^n |Ab_i|^2 - \sum_{i=1}^{2k} |Ab_i|^2 \\ &\stackrel{\text{Parseval's id.}}{=} \|A\|_F^2 - \sum_{i=1}^{2k} |Ab_i|^2 \end{aligned}$$

On the other hand, we have

$$\begin{aligned}\|A - B_{2k}\|_F^2 &\stackrel{\text{E-Y-M theorem}}{=} \sum_{i=k+1}^n \sigma_i^2 \\ &\stackrel{\text{Frob. norm def.}}{=} \|A\|_F^2 - \|A_k\|_F^2.\end{aligned}$$

Now we consider

$$\begin{aligned}\|A - B_{2k}\|_F^2 - \|A - A_k\|_F^2 &= \|A\|_F^2 - \sum_{i=1}^{2k} |Ab_i|^2 - (\|A\|_F^2 - \|A_k\|_F^2) \\ &= \|A_k\|_F^2 - \sum_{i=1}^{2k} |Ab_i|^2\end{aligned}$$

That is equivalent to

$$\|A - B_{2k}\|_F^2 = \|A - A_k\|_F^2 + (\|A_k\|_F^2 - \sum_{i=1}^{2k} |Ab_i|^2)$$

For the next step, we show

$$(1 + \varepsilon) \sum_{i=1}^{2k} |Ab_i|^2 \geq \sum_{i=1}^{2k} \lambda_i^2.$$

$$\sum_{i=1}^{2k} \lambda_i^2 \stackrel{|Bb_i|=\lambda_i}{=} \sum_{i=1}^{2k} |Bb_i|^2 \stackrel{\text{subst. B}}{=} \sum_{i=1}^{2k} \left| \sqrt{\frac{n}{\ell}} R^T(Ab_i) \right|^2 = \sum_{i=1}^{2k} \frac{n}{\ell} |R^T(Ab_i)|^2$$

Now from the Random projection theorem for very large ℓ we have for each i

$$\frac{n}{\ell} |R^T(Ab_i)|^2 \leq (1 + \varepsilon) |Ab_i|^2$$

with high probability.

Hence with a high probability

$$(1 + \varepsilon) \sum_{i=1}^{2k} |Ab_i|^2 \geq \sum_{i=1}^{2k} \lambda_i^2.$$

Now we have

$$\sum_{i=1}^{2k} |Ab_i|^2 \geq \frac{1}{(1 + \varepsilon)} \sum_{i=1}^{2k} \lambda_i^2 \stackrel{\text{Cor.11}}{\geq} \frac{(1 - \varepsilon)}{(1 + \varepsilon)} \|A_k\|_F^2 \geq (1 - 2\varepsilon) \|A_k\|_F^2$$

I.e.

$$\sum_{i=1}^{2k} |Ab_i|^2 \geq (1 - 2\varepsilon) \|A_k\|_F^2$$

Now we substitute this result in [BUILD REFFFFFFF] getting

$$\begin{aligned}\|A - B_{2k}\|_F^2 &\leq \|A - A_k\|_F^2 + \|A_k\|_F^2 - (1 - 2\varepsilon) \|A_k\|_F^2 \\ &\iff \|A - B_{2k}\|_F^2 \leq \|A - A_k\|_F^2 + 2\varepsilon \|A_k\|_F^2\end{aligned}$$

Due to the definition of Frobenius norm, we have $\|A\|_F^2 \geq \|A_k\|_F^2$. This inequality leads us to the conclusive result of our proof

$$\|A - B_{2k}\|_F^2 \leq \|A - A_k\|_F^2 + 2\varepsilon\|A\|_F^2.$$

What are the computational savings achieved by the two-step method? Let A be an $n \times m$ matrix. The time to compute LSI is $O(mnc)$ if A is sparse with about c nonzero entries per column (i.e., c is the average number of terms in a document). The time needed to compute the random projection to l dimensions is $O(mcl)$. After the projection, the time to compute LSI is $O(ml^2)$. So the total time is $O(ml(1+c))$. To obtain an ε approximation we need l to be $O(\sqrt{\log n})$. Thus the running time of the two-step method is asymptotically superior: $O(m(\sqrt{\log n} + c \log n))$ compared to $O(mnc)$. \square

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.

Let $A \in \mathbb{R}^{n \times m}$ be a matrix and b_1, \dots, b_n an orthonormal set of vectors where $b_i \in \mathbb{R}^m$. F is the Frobenius norm of a matrix. Why is the following correct?

$$\|A\|_F^2 = \sum_{i=1}^n |Ab_i|^2$$

References