Latent Semantic Indexing

Seminar "Theoretical Topics in Data Science"

Vahe Eminyan

November 11, 2023

RWTH Aachen University

This seminar paper gives an analysis of the work of Papadimitriou et al. "Latent Semantic Indexing: A Probabilistic Analysis" [10].

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 [2, 8, 1]. 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 delve into both aforementioned aspects, with a specific emphasis on the second question.

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 Related Work

Hofmann et al. analyze the LSI from a different probabilistic point of view comparing SVD to mixture decomposition using expectation maximization algorithm [4]. Despite analyzing the LSI's good performance, Frieze et al. give another method to speed up LSI, introducing an approach that includes Fast-Monte-Carlo algorithms [3].

3 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 ith term of the entire term space. The function can, for example, be the frequency of the ith 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

$$A = UDV^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T$$

Where $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r$ are the eigenvalues of the matrix AA^T (called singular values of A), $D = diag(\sigma_1, ..., \sigma_r) \in \mathbb{R}^{r \times r}$. $U \in \mathbb{R}^{n \times r}$ 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 1. Every matrix can be represented in this form [11].

$$A = \begin{bmatrix} | & | & & | \\ u_1 & u_2 & \cdots & u_r \\ | & | & & | \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r \end{bmatrix} \begin{bmatrix} | & | & & | \\ v_1 & v_2 & \cdots & v_r \\ | & | & & | \end{bmatrix}^T \tag{1}$$

Now 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. U can be interpreted as "term-topic similarity" matrix, where each row represents how the given word is related to each of the r-topics. V can be interpreted as "document-topic similarity" matrix, where each row represents how the given document relates to each r-topics. Each diagonal element of the matrix D shows the relevance of each topic ordered in decreasing order (from more relevant to less relevant).

Rank-k LSI keeps only k-largest singular values of the matrix considering only the first k columns of matrices U and V(k defines the top k most important topics/concepts of the term-document matrix). 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 rank k approximation of matrix A denoted as A_k is the matrix that minimizes the Frobenius norm of $A - A_k$. Formally:

Theorem 1. *[6] 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 show why it finds queries and documents that are semantically related.

4 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. A topic is a probability distribution on the universe of all terms 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 3. 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 4. A corpus model $\mathscr{C}$ is a quadruple $\mathscr{C} = (U, \mathscr{T}, \mathscr{S}, D)$, where U is the universe of terms, $\mathscr{T}$ is a set of topics, and $\mathscr{S}$ is a set of styles, and D a probability distribution on $\hat{\mathscr{T}} \times \hat{\mathscr{S}} \times \mathbb{Z}^+$, where by $\hat{\mathscr{T}}$ we denote the set of all convex combinations of topics in $\mathscr{T}$, by $\hat{\mathscr{S}}$ the set of all convex combinations of styles in $\mathscr{S}$, 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 $\hat{T} \in \mathscr{T}$, $\hat{S} \in \mathscr{S}$, and a fixed length ℓ of the document. Then we sample the terms of the document ℓ times according to $\hat{T}$ [10].

5 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 5. A corpus model $\mathscr{C}$ is pure if each document involves a single topic.

Definition 6. A corpus model $\mathscr{C}$ is ε-separable ($\varepsilon \in [0, 1)$), if a set of terms U_T is associated with each topic $T \in \mathscr{T}$ 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 $\mathscr{C} = (U, \mathscr{T}, D)$ be a pure and style-free corpus model and let $k = |\mathscr{T}|$ denote the number of topics in $\mathscr{C}$. Since $\mathscr{C}$ is pure, each document generated from $\mathscr{C}$ 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 $\mathscr{C}$ 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 7. 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 8. *Let $\mathscr{C}$ 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 $\mathscr{C}$. Then, the rank-k LSI is $O(\varepsilon)$-skewed on C with probability $1 - O(m^{-1})$.*

The proof of this theorem can be found in the original paper [10].

6 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 [7] 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 onto ℓ dimensions, where ℓ is a small value greater than k, on the initial corpus. This process, with high probability, yields a significantly smaller representation that maintains close proximity to the original corpus in both distances and angles.

$$B = \sqrt{\frac{n}{l}} \cdot \begin{bmatrix} | & | & & | \\ r_1 & r_2 & \cdots & r_\ell \\ | & | & & | \end{bmatrix}^T \cdot A \tag{2}$$

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 $R \in \mathbb{R}^{n \times l}$ 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, see the visualization in 2. The SVD representations of A and B are denoted as follows:

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

We also introduce two lemmas, a corollary and a theorem, which we use in our proof of the main theorem 13.

Lemma 9. *Let ε be an arbitrary positive constant. If $\ell \geq c((\log n)/\varepsilon^2)$ for a sufficiently large constant c then, for $p = 1, ...t$*

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

Corollary 10.

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

Lemma 11.

$$\|A - A_k\|_F^2 = \sum_{i=k+1}^{n} \sigma_i^2.$$

Proof.

$$\|A - A_k\|_F^2 = \| \sum_{i=k+1}^{n} \sigma_i u_i v_i^T \|_F^2 = \sum_{i=k+1}^{n} \sigma_i^2$$

The second equality holds since the squared Frobenius norm of every matrix can be written as sum of squares of its singular values [9]. □

Theorem 12. *Parsevals identity [5]: Let $b_1, ..., b_n$ be an orthonormal basis for a space S. Then for each $s \in S$, $|s|^2 = \sum_{i=1}^{n} (s b_i)^2$.*

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 13.

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

where $\varepsilon \in (0, 0.5)$

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

$$A = \sum_{i=1}^{n} \sigma_i u_i v_i^T, \qquad A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T, \qquad B = \sum_{i=1}^{t} \lambda_i a_i b_i^T, \qquad B_{2k} = A\sum_{i=1}^{2k} b_i b_i^T.$$

$b_1, ..., b_n$ is an orthonormal set of vectors spanning the row space of A (because R is an orthonormal matrix, resulting orthonormal projection) 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 $i = 1, ..., 2k$, because $b_i^T b_i = 1$ we have

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

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

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

Hence,

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

On the other hand, we have

$$\|A - A_k\|_F^2 \overset{\text{Lemma 11}}{=} \sum_{i=k+1}^{n} \sigma_i^2 \overset{\text{Frob. form. [9]}}{=} \|A\|_F^2 - \|A_k\|_F^2.$$

Now we consider

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

That is equivalent to

$$\|A - B_{2k}\|_F^2 = \|A - A_k\|_F^2 + (\|A_k\|_F^2 - \sum_{i=1}^{2k} |Ab_i|^2) \tag{3}$$

For the next step, we show

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

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

Now from the Johnson-Lindenstrauss lemma [7] for very large $\ell \in \Omega((\log n)/\varepsilon^2)$ 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

$$\sum_{i=1}^{2k} |Ab_i|^2 \geq \frac{1}{(1+\varepsilon)} \sum_{i=1}^{2k} \lambda_i^2 \overset{\text{Cor. 10}}{\geq} \frac{(1-\varepsilon)}{(1+\varepsilon)} \|A_k\|_F^2 \geq (1-2\varepsilon)\|A_k\|_F^2$$

I.e.

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

Now we substitute this result in equation 3 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 formulation of Frobenius norm as in 11, we have $\|A\|_F^2 \geq \|A_k\|_F^2$. This inequality leads us to the conclusive result of our proof

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

□

The two-step method offers significant computational savings compared to traditional Latent Semantic Indexing (LSI). Consider a matrix A of size $n \times m$. If A is sparse, containing approximately c nonzero entries per column (where c represents the average number of terms in a document), the time complexity to compute Latent Semantic Indexing (LSI) is $O(mnc)$. The time complexity for computing the random projection to ℓ dimensions is $O(mc\ell)$. After the projection, we compute LSI. Its computation is in $O(m\ell^2)$. Hence the total time is $O(m\ell(\ell + c))$. For an ε approximation ℓ has to be in $O((\log n)/\varepsilon^2)$. So, the two-step method outperforms the single-step method in terms of runtime. The time complexity is $O(m(\log^2 n + c \log n))$ for the two-step method, whereas it is $O(mnc)$ for the single-step method.

7 Summary and Conclusion

In our seminar paper, we gave the most relevant information of the scientific paper from Papadimitriou et al. [10] introducing the theorem 8 which shows the LSI's good performance assuming some constraints on the given term-document matrix. These assumptions are sometimes not given in a real-world scenario. Hence the authors hope that this work will serve as a motivation for further work where these constraints on the term-document matrix are not given.

Furthermore, we considered the LSI by random projection a two-step LSI which reduces the computational time. We showed the central theorem which argued that the two-step LSI is close to the single-step LSI with high probability 13 also giving a detailed proof of the theorem.

References

[1] Susan T Dumais, Todd A Letsche, Michael L Littman, and Thomas K Landauer. Automatic cross-language retrieval using latent semantic indexing. In *AAAI spring symposium on cross-language text and speech retrieval*, volume 15, page 21. Stanford University Stanford, CA, USA, 1997.

[2] P. W. Foltz. Using latent semantic indexing for information filtering. In *Proceedings of the ACM SIGOIS and IEEE CS TC-OA Conference on Office Information Systems*, COCS '90, page 40–47, New York, NY, USA, 1990. Association for Computing Machinery. doi:10.1145/91474.91486.

[3] Alan Frieze, Ravindran Kannan, and Santosh Vempala. Fast monte-carlo algorithms for finding low-rank approximations. volume 51, pages 370–378, 12 1998. doi:10.1109/SFCS.1998.743487.

[4] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, page 50–57, New York, NY, USA, 1999. Association for Computing Machinery. doi:10.1145/312624.312649.

[5] Leslie Hogben, editor. *Handbook of Linear Algebra*. Chapman and Hall/CRC, 2nd edition, 2013. https://doi.org/10.1201/b16113.

[6] C. Reinsch J. H. Wilkinson. *Handbook for Automatic Computation*. Springer Berlin, Heidelberg, volume ii: linear algebra edition, 1971.

[7] William B Johnson. Extensions of lipshitz mapping into hilbert space. In *Conference modern analysis and probability, 1984*, pages 189–206, 1984.

[8] Todd A. Letsche and Michael W. Berry. Large-scale information retrieval with latent semantic indexing. *Information Sciences*, 100(1):105–137, 1997. URL: https://www.sciencedirect.com/science/article/pii/S0020025597000443, doi:https://doi.org/10.1016/S0020-0255(97)00044-3.

[9] Changxue Ma, Y. Kamp, and L.F. Willems. A frobenius norm approach to glottal closure detection from the speech signal. *IEEE Transactions on Speech and Audio Processing*, 2(2):258–265, 1994. doi:10.1109/89.279274.

[10] 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:https://doi.org/10.1006/jcss.2000.1711.

[11] Gilbert Strang. *Linear Algebra and Its Applications*. Cengage Learning, 4th edition edition, 2005.