latent semantic indexing twenty years later

Christos H. Papadimitriou
Columbia University



"Latent Semantic Indexing: a probabilistic analysis" by CHP, Prabhakar Raghavan, Hisao Tamaki, Santosh Vempala PODS 1999

the summer of 1998

- The Internet
- The web
- The web search problem
- Lycos, Alta Vista, Inktomi, Yahoo, Google, Overture…
- The Theory Group at IBM Almaden
- Hubs and authorities, communities, the bow-tie web, …

Latent semantic indexing

- Gerald Salton, 1970s and 1980s

  "the corpus is a matrix"

- Deerwester, Dumais, Furnas, Landauer 1990

  "then apply SVD" → LSI

- 1990 - 1998: LSI is very successful in practice!

TCS in the 20th century: the three missions

1. Understand through math the power and limitations of computers
2. Guide computing practice by discovering through math the right way to do things
3. Annoy practitioners by proving that what they are already doing is fine

Why is LSI so successful?

- SVD projects the matrix to the subspace of the principal directions (= “virtual words”)
- (those with the highest “eigenvalues”)
- It is the “minimally distorting projection”
- The resulting representation of the corpus works better (eg, NN search seems to yield better results...)
- ***Why?***

Our suspicion

- Can it be that LSI projection identifies the "topics" of the document?
- Suppose that topics (politics, sports, science, art, commerce) have each its own word distribution
- Every document is a mixture of topics
- This suggests a generative model
- Does LSI identify the topics of the document?

Remember the historical context: summer of 1998

- Extracting insight from soulless data was a strange and mysterious phenomenon, in need of some explanation
- Machine learning was something Les Valiant does
- Less than 5% of the ~200 papers in NIPS were about neural nets

From the introduction

"…We would like to prove a theorem stating essentially that if the corpus is a reasonably focused collection of meaningfully correlated documents, then LSI performs well. The problem is to define these terms so that (1) there is a reasonably close correspondence with what they mean intuitively and in practice, and (2) the theorem can be proved."

Our theorems

- Under strong assumptions of separation of distributions and of mixtures of topics, LSI does identify the main topics of the document, whp.
- Also, random projection combines well with LSI, and saves much work, whp
- And in experiments with this generative model, LSI + RP works much better than we can prove…

What happened

- Paper was presented at PODS 1999 in Philly…
- …and was selected for the special volume
- Essentially the same generative model was formulated and treated as a machine learning problem in AI
- T. Hofmann 1999: pLSI,
- D. Blei, M. Jordan, A. Ng 2003: LDA

What else has happened since 1999

- Explosive growth and success of machine learning and neural nets
- Web search engine companies are at the forefront of this revolution
- A dearth of ex post math explanations
- …and the web is no longer the promise of futuristic utopia it once was…


Thank You!
PASAJE GALVEZ

"Latent Semantic Indexing: a probabilistic analysis" by CHP, Prabhakar Raghavan, Hisao Tamaki, Santosh Vempala PODS 1999