Iterative Residual Rescaling: An analysis and generalization of Latent Semantic Indexing

Lillian Lee
Cornell University
http://www.cs.cornell.edu/home/llee

Joint work with Rie Kubota Ando

SIGIR 2001
The Document Representation Problem

**Goal:** Find a representation that succinctly describes the “meaning” of a “document” ...

... or in which we at least can determine if two “documents” have “similar” “meanings”, *without human labelings.*

- information retrieval
- multi-document summarization
- topic spotting
- creating/organizing knowledge resources
The Vector Space Model (VSM)

Documents:

Term–document matrix $D$:

Vector space:

large angle = small cosine
Problems: Synonymy & Polysemy

Documents:

Term-document matrix $D$:

Vector space:

large cosine, but not truly related
Approach: Subspace Projection

Given a term-document matrix $D$, project the document vectors into a different subspace so that vector cosines more accurately represent semantic similarity.

In a lower dimensional space, synonym vectors may not be orthogonal.

Latent Semantic Indexing [Deerwester, Dumais, Furnas, Landauer, Harshman 1990] seeks to uncover such hidden semantic relations through projection methods.

Talk Outline

- Introduction: Latent Semantic Indexing (LSI)
- A new analysis: relating LSI's potential to the uniformity of the underlying topic-document distribution [Ando+Lee 2001]
- Experimental results
Introduction to LSI
Singular Value Decomposition

The SVD is the matrix factorization underlying LSI.

Let the $m \times n$ term-document matrix $D$ have rank $r$.

$$D = U \times \Sigma \times V^T$$

$u_i$: left singular vectors; form a basis for range($D$)

$\sigma_i$: singular values (assume in sorted order); all positive

(Each $u_i$ is an eigenvector of $DD^T$ with eigenvalue $\sigma_i^2$)
SVD: Geometric View

Recall:

More formally, find \( r = \text{rank}(D) \) vectors such that

\[
\mathbf{u} = \arg \max_{\mathbf{x} : |\mathbf{x}| = 1} \sum_{j=1}^{n} |r_j|^2 \cos^2(\angle(\mathbf{x}, r_j)) \quad ("\text{weighted average}")
\]
Latent Semantic Indexing

LSI projects $D$ into the $h$-dimensional subspace spanned by $u_1, \ldots, u_h$.

$D' = U \times \Sigma' \times V^T$

$\begin{bmatrix}
  d_1 & d_2 & \cdots & d_n
\end{bmatrix}
\begin{bmatrix}
  u_1 \\
  u_2 \\
  \vdots \\
  u_r
\end{bmatrix}$

**Theorem**: This is the optimum (in two-norm) rank-$h$ approximation to $D$. (Note that it selects the $h$ basis vectors that maximize projections.)
LSI (continued)

Recall: LSI computes the optimum rank-$h$ approximation to $D$.

But this does not mean LSI does the best job at representing document relationships – just the best job at being close to $D$.


(See e.g. [Dumais+al 1998])

We desire an analysis based on the underlying semantic relationships.
Analyzing LSI
Topic Model

For a given set of $n$ documents, we assume there exists the following unknown quantities:
- a set of $k < n$ underlying topics
- (normalized) document-topic relevance scores

These define the hidden true topic-based document similarities:

$$\text{sim}(\text{doc, doc}') = \sum_{\text{topics } t} \text{rel}(\text{doc, } t) \times \text{rel}(\text{doc}', t)$$

and we desire a subspace in which vector cosines approximate these true similarities closely.
Subspace Projections

Let \( \mathcal{X}^{opt} \) be the subspace with \textit{minimum} similarity error (and dimensionality) where \( \text{error}(\mathcal{X}) = \| [\text{sim}(\text{doc}_i, \text{doc}_j)] - \frac{d_i^X \cdot d_j^X}{\| d_i^X \| \| d_j^X \|} \|_2 \)

How close is \( \mathcal{X}^{LSI} \) to \( \mathcal{X}^{opt} \)? Let’s define some useful quantities …
Dominance and Non-Uniformity

The (hidden) dominance of a topic in the document collection is defined as:

\[ \text{Dom}(t) = \sqrt{\sum_{\text{doc}} \text{rel(doc, } t\text{)}^2} \]

We assume a dominance ordering on the topics, most dominant first.

Intuitively, less dominant topics risk being “lost”.

\[ \text{Dom} \gg \text{Dom} \gg \text{Dom} \]

\[ \text{non-uniformity} = \frac{\text{Dom}}{\text{Dom}} \text{ is high} \]

\[ \text{non-uniformity is low} \]
Document Sharing and Difficulty

The (hidden) degree to which topics share documents is defined as:

\[
\text{DocSharing} = \sqrt{\sum_{t \neq t'} \left( \sum_{\text{doc}} \text{rel}(\text{doc}, t)\text{rel}(\text{doc}, t') \right)^2}
\]

more document sharing among topics

less document sharing (same dominances)

Intuitively, when document sharing is high, distinguishing between topics is difficult. ([Papadimitriou+al 1997] assume low document sharing.)
Structure of Main Result

The distance between $\mathcal{X}^{LSI}$ and $\mathcal{X}^{opt}$ can be bounded by a function of:

- $\text{error}(\mathcal{X}^{VSM})$ and $\text{error}(\mathcal{X}^{opt})$,
- the amount of document sharing between topics, and
- the non-uniformity of the topic-document distribution, as measured by a ratio of topic dominances.

assuming that $\text{error}(\mathcal{X}^{VSM})$ doesn’t swamp certain topic dominances.

The proof relies on:

1) a \textit{subspace perturbation theorem} [Stewart 1973, Davis+Kahan 1970] relating subspace distances to certain singular values, and

2) \textit{sensitivity} theorems relating certain singular values to topic dominances.
Related Work


\[
\begin{array}{c}
\text{CHOOSE} \\
\text{subspace } X
\end{array}
\quad
\begin{array}{c}
\text{GIVEN} \\
\text{term-doc matrix } D
\end{array}
\quad
\begin{array}{c}
\text{HIDDEN} \\
\text{true similarities}
\end{array}
\]

\textit{orthogonal projection}

similarities (cosine) in \( X \)

\[ \text{true similarities} \]

\[ \text{orthogonal projection} \]

\[ \text{similarities (cosine) in } X \]

\[ \text{true similarities} \]

Cf. our framework:

\[
\begin{array}{c}
\text{CHOOSE} \\
\text{subspace } X
\end{array}
\quad
\begin{array}{c}
\text{GIVEN} \\
\text{term-doc matrix } D
\end{array}
\quad
\begin{array}{c}
\text{HIDDEN} \\
\text{topic-document relevances}
\end{array}
\]

\textit{orthogonal projection}

\[ \text{true similarities} \]

\[ \text{true similarities} \]

\[ \text{true similarities} \]

The Iterative Residual Rescaling (IRR) Algorithm
Non-uniformity: Geometric Interpretation

LSI finds a sequence of $h$ basis vectors such that

$$u = \arg \max_{x: |x|=1} \sum_{j=1}^{n} |r_j|^2 \cos^2(\angle(x, r_j))$$

("weighted average")

Choose direction $u$ maximizing projections

Compute residuals

Repeat to get next $u$ (orthogonal to previous $u_i$'s)

dominant topics bias the choice
IRR: First Version

\[ u = \arg \max_{x: |x| = 1} \sum_{j=1}^{n} |r_j|^2 \cos^2(\angle(x, r_j)) \] ("weighted average")

*Compensate* for non-uniformity by rescaling the residuals by the \( q \)th power of their length at each iteration. [Ando 2000]

Choose direction \( u \) maximizing projections

Compute residuals

Rescale residuals (relative diffs rise)

Repeat to get next \( u \) (orthogonal to previous \( u_i \)'s)

Good results, but how do we pick the scaling factor \( q \)?

We need a *principled* way to choose amount of re-scaling.
Scaling Factor Determination

Consider the following function of non-uniformity: \( \sum_t \left( \frac{\text{Dom}(t)^2}{n} \right)^2 \)

- one giant topic \( \rightarrow 1 \)
- \( k \) same-size topics with no document sharing \( \rightarrow \frac{1}{k} \)

We’d like to set the scaling factor \( q \) to this quantity to compensate for non-uniformity ...

but we don’t know it!

We can roughly approximate it in our model by
\[
\sum_{d_i,d_j} \cos^2(\angle(d_i,d_j))/n^2.
\]
(coarse assumptions: small input error, single-topic documents)

We set \( q \) to a linear function of this approximation.
Experiments
Experimental Framework: Data

We used TREC documents, with topic labels as validation. (Stop-words removed; no term weighting; only single-topic documents (no topic sharing) to facilitate scoring).

**Controlled distributions**: we artificially altered topic dominances to study their effects on LSI and IRR’s performance

- For a set of $k$ topics, for a sequence of increasingly non-uniform distributions, ten 50-document sets were selected randomly for each.

**Uncontrolled distributions**: we simulated retrieval results.

- For each keyword in a randomly-chosen set of 15, all documents containing that keyword were selected to create a document set.
Evaluation Metrics

Kappa average precision: degree to which same-topic document pairs have high similarity scores, corrected for chance

Clustering score: degree to which a clustering has “pure” clusters but preserves topic integrity [cf. Slonim and Tishby 2000]

We record the floor and ceiling results over 6 clustering algorithms. A high-quality subspace should enable good results for many clustering algorithms.

[To simplify presentation, we do not discuss dimensionality selection issues]
(Switch to slides on experimental results now)