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)

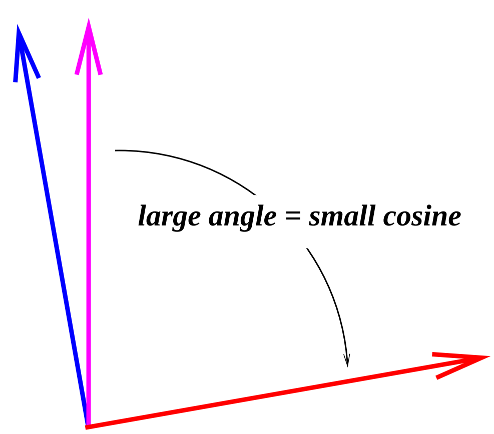
Documents:

Document 1	Document 2	Document 3
car	car	Chomsky
engine	emissions	corpus
hood	hood	noun
tires	make	parsing
truck	model	tagging
trunk	trunk	wonderful

Term−document matrix D:

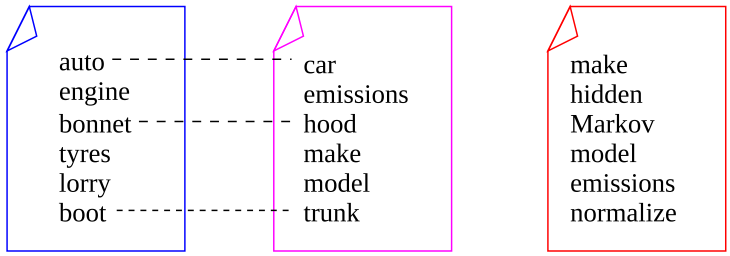
1	1	0	car
0	0	1	Chomsky
0	0	1	corpus
0	1	0	emissions
1	0	0	engine
1	1	0	hood
0	1	0	make
0	1	0	model
0	0	1	noun
0	0	1	parsing
0	0	1	tagging
1	0	0	tires
1	0	0	truck
1	1	0	trunk
0	0	1	wonderful

Vector space:



Problems: Synonymy & Polysemy

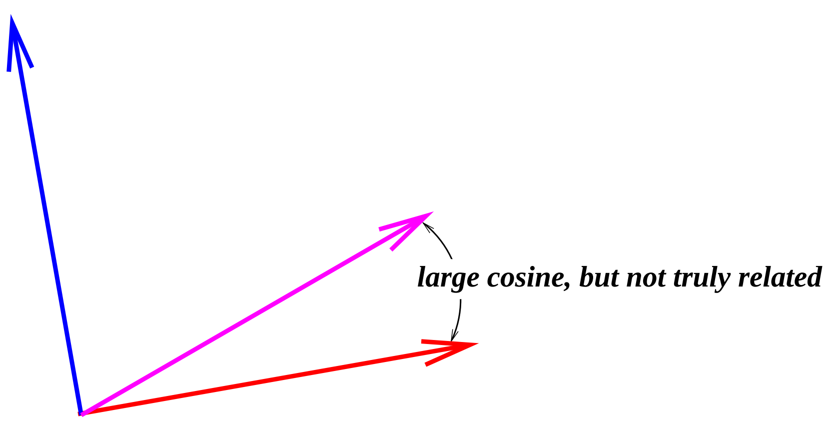
Documents:



Term-document matrix D:

1	0	0	auto
1	0	0	bonnet
1	0	0	boot
0	1	0	car
0	1	1	emissions
1	0	0	engine
0	0	1	hidden
0	1	0	hood
1	0	0	lorry
0	1	1	make
0	0	1	Markov
0	1	1	model
0	0	1	normalize
0	0	0	tires
0	1	0	trunk
1	0	0	tyres

Vector space:



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.

Applications (a sampling): [Dumais 1991, 1993, 1994, 1995], [Landauer+Littman 1990], [Foltz 1990, 1996], [Foltz+Dumais 1992], [Dumais+Nielsen 1992], [Foltz+al 1996, 1998a, 1998b], [Landauer+al 1997, 1998], [Schütze+Silverstein 1997], [Soboroff+al 1998], [Wolfe+al 1998], [Weimer-Hastings, 1999], [Jiang+al 1999b], [Kurimo 2000] [Weimer-Hastings+al, 1999], [Schone+Jurafsky 2000, 2001]

Talk Outline

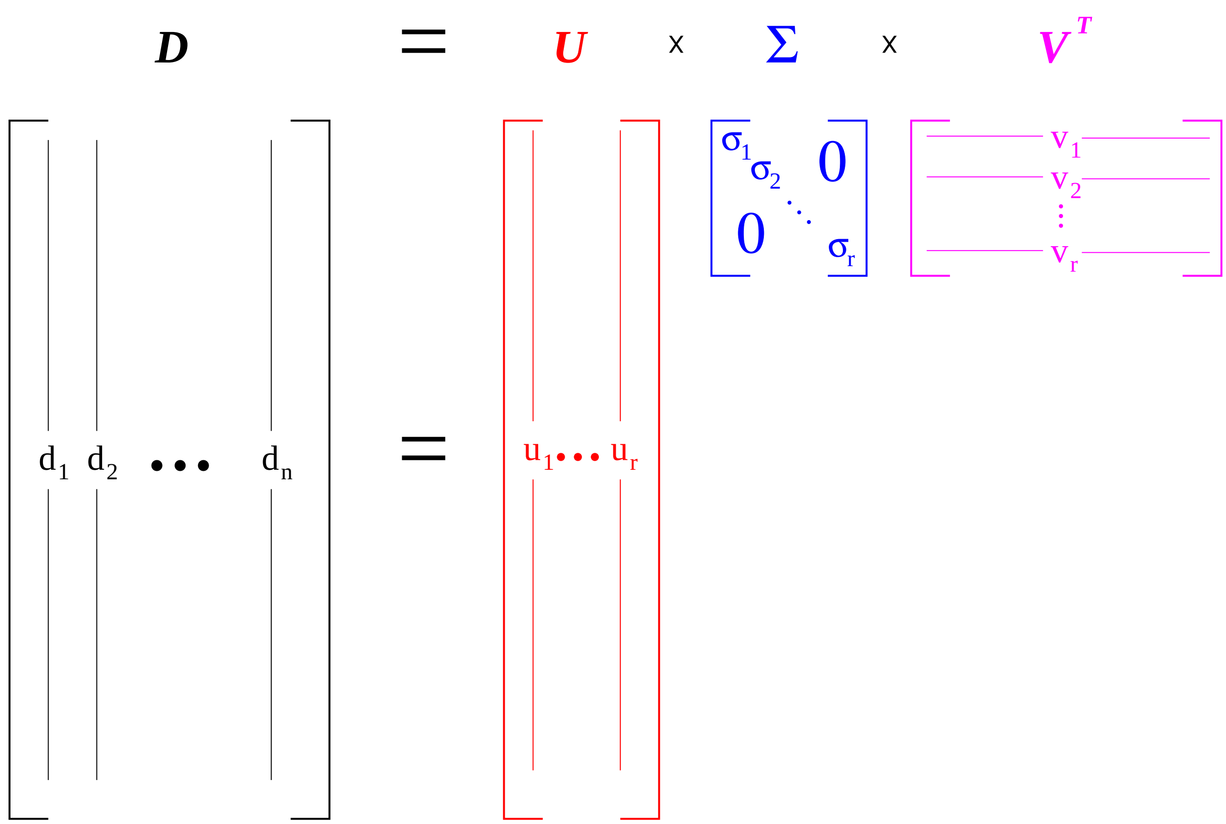
- Introduction: Latent Semantic Indexing (LSI)
- A new analysis: relating LSI's potential to the uniformity of the underlying topic-document distribution [Ando+Lee 2001]
- A new algorithm: Iterative Residual Rescaling automatically compensates for non-uniformity [Ando 2000; Ando+Lee 2001]
- Experimental 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.

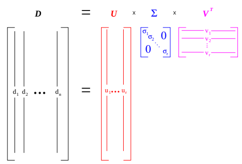


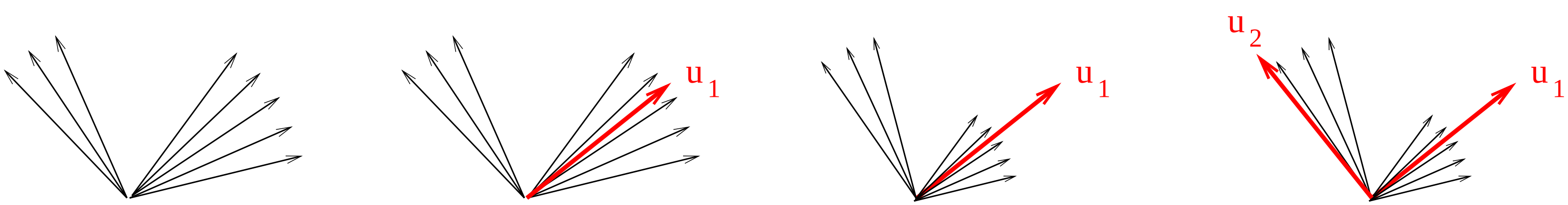
u_i: left singular vectors; form a basis for range(D)

σ_i: singular values (assume in sorted order); all positive

(Each u_i is an *eigenvector* of DD^T with eigenvalue σ_i^2)

SVD: Geometric View

Recall: 



Start with document vectors

Choose direction u maximizing projections (σ : "sizes" of max. projection)

Compute residuals (subtract projections)

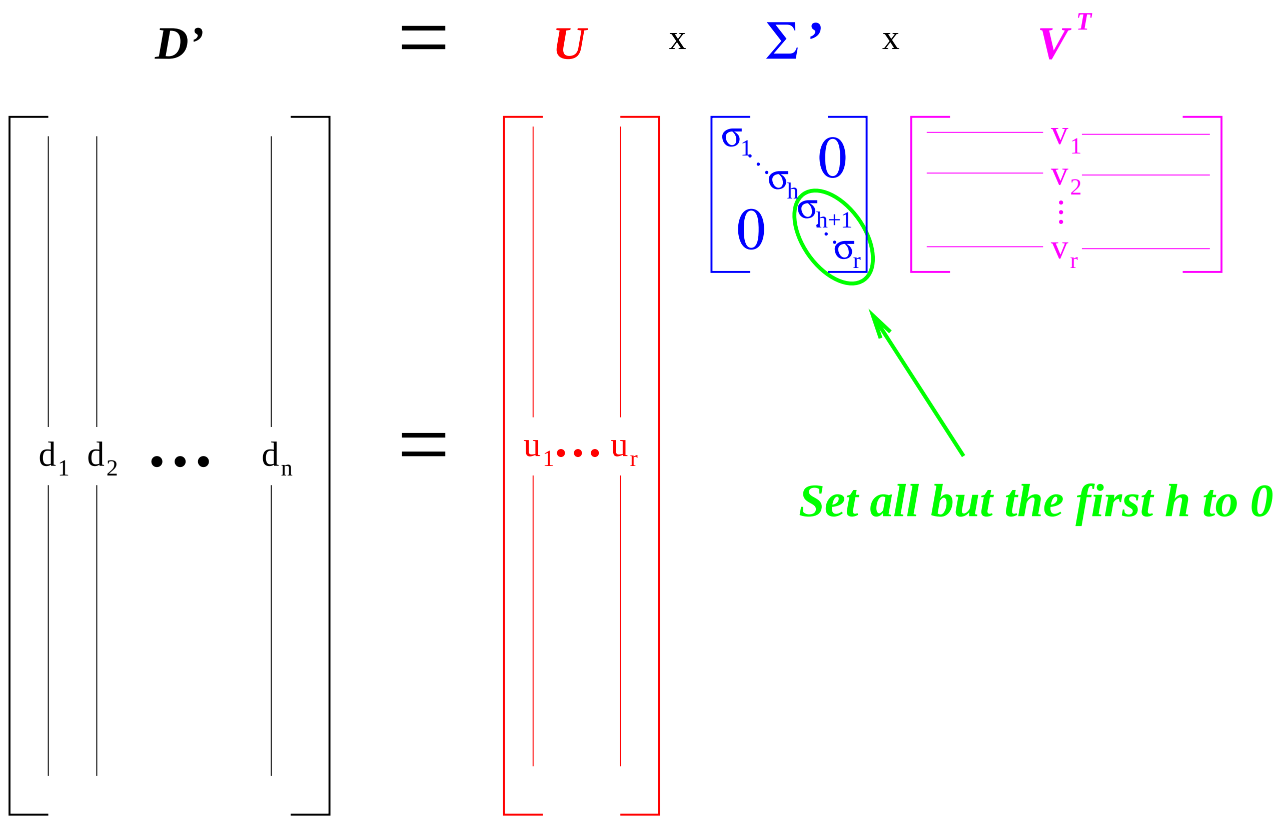
Repeat to get next u (orthogonal to previous u_i's)

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

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

Latent Semantic Indexing

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



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.

> "Whether [LSI] is superior in practical situations with general collections remains to be verified." Baeza-Yates and Ribeiro-Neto, *Modern Information Retrieval*, 1999.

(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:

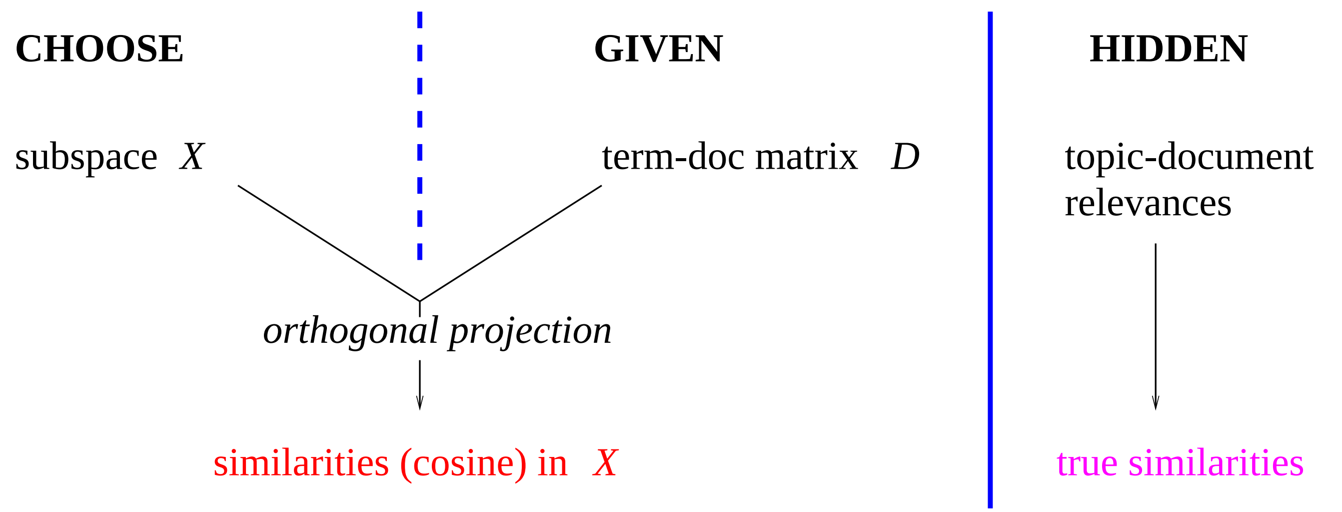
- a set of $k < n$ underlying topics
- (normalized) document-topic relevance scores

These define the hidden true topic-based document similarities:

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

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

Subspace Projections



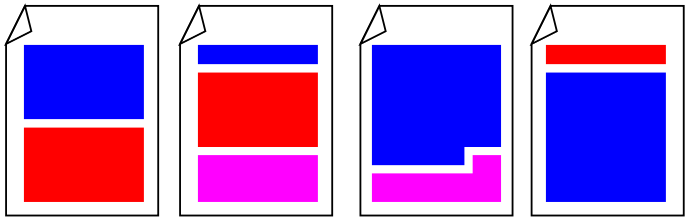
Let $\mathcal{X}^{opt}$ be the subspace with *minimum* similarity error (and dimensionality) where $\mathrm{error}(\mathcal{X}) = ||[\mathrm{sim}(\mathsf{doc}_i, \mathsf{doc}_j)] - \underline{d_i^{\mathcal{X}} \cdot d_j^{\mathcal{X}}}||_2$

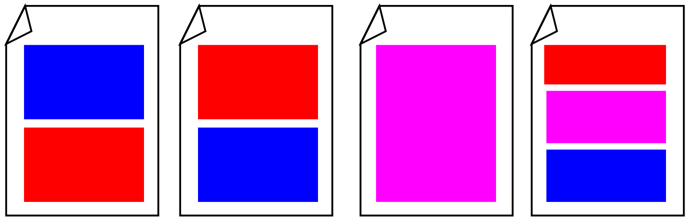
How close is $\mathcal{X}^{LSI}$ to $\mathcal{X}^{opt}$? Let's define some useful quantities ...

Dominance and Non-Uniformity

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

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





Dom >> Dom >> Dom

non-uniformity = Dom / Dom is high

Dom = Dom = Dom

non-uniformity is low

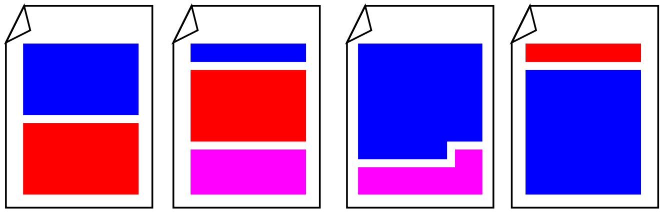
We assume a dominance ordering on the topics, most dominant first.

Intuitively, less dominant topics risk being “lost”.

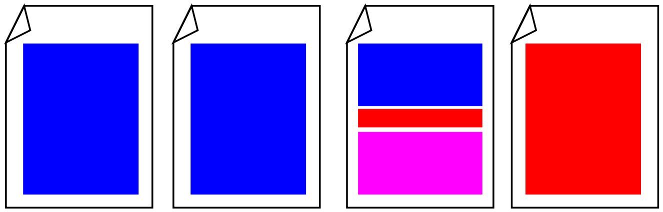
Document Sharing and Difficulty

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:

- $\mathrm{error}(\mathcal{X}^{VSM})$ and $\mathrm{error}(\mathcal{X}^{opt})$,
- 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 $\mathrm{error}(\mathcal{X}^{VSM})$ doesn't swamp certain topic dominances.

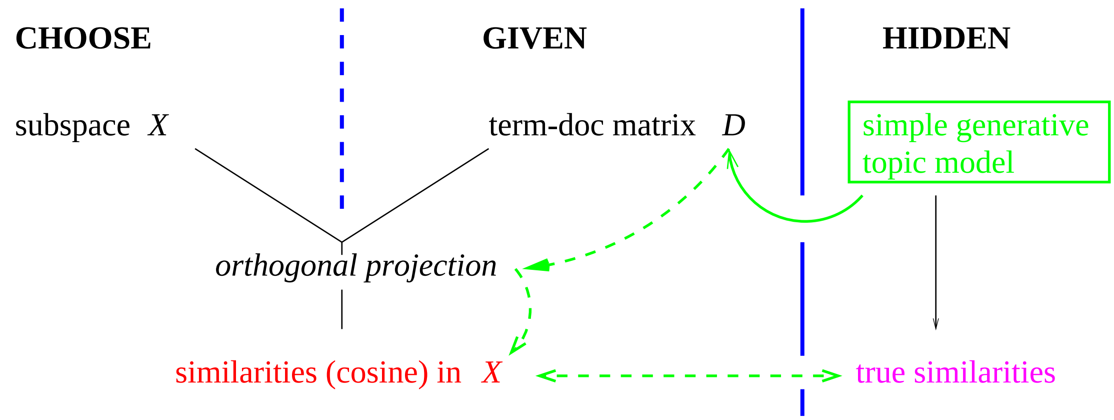
The proof relies on:

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

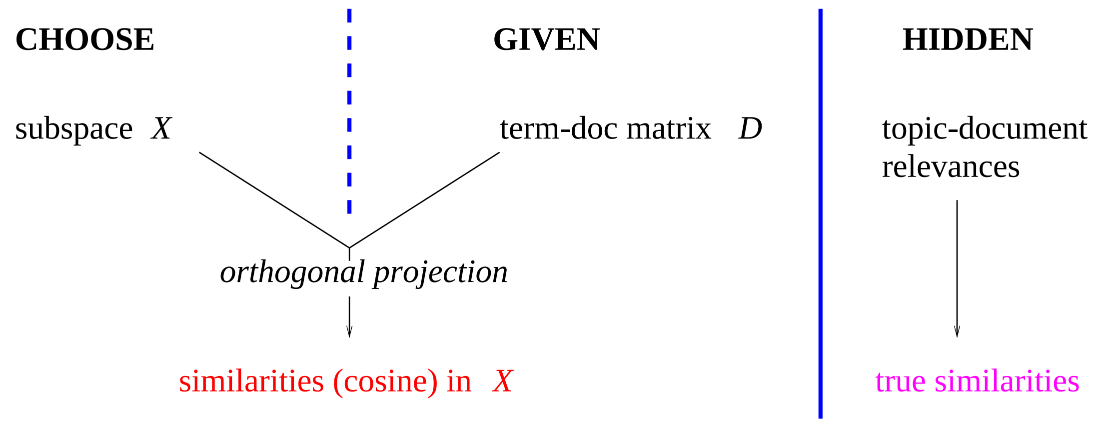
2) *sensitivity* theorems relating certain singular values to topic dominances.

Related Work

[Papadimitriou+al 1997, Azar+al 2001, Story 1996, Ding 1999] etc. assume a *generative* model in which LSI "works"



Cf. our framework:



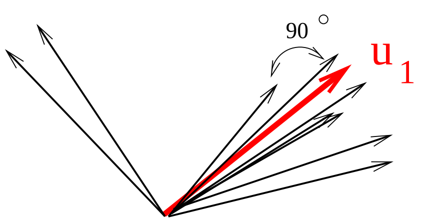
(cf. [Bartell+Cottrell+Belew 1992; 1995, Isbell+Viola 1998])

The Iterative Residual Rescaling (IRR) Algorithm

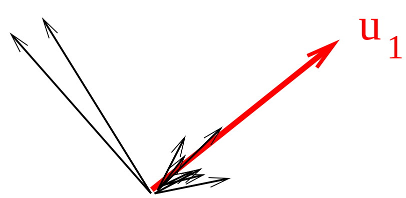
Non-uniformity: Geometric Interpretation

LSI finds a sequence of h basis vectors such that

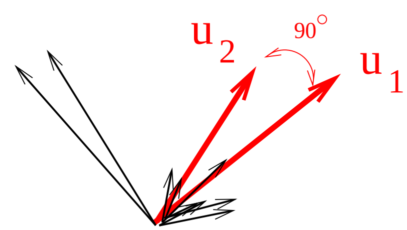
$u = \arg\max_{x:|x|=1} \sum_{j=1}^{n} |r_j|^2 \cos^2(\angle(x, r_j))$ ("weighted average")



Choose direction u maximizing projections



Compute residuals



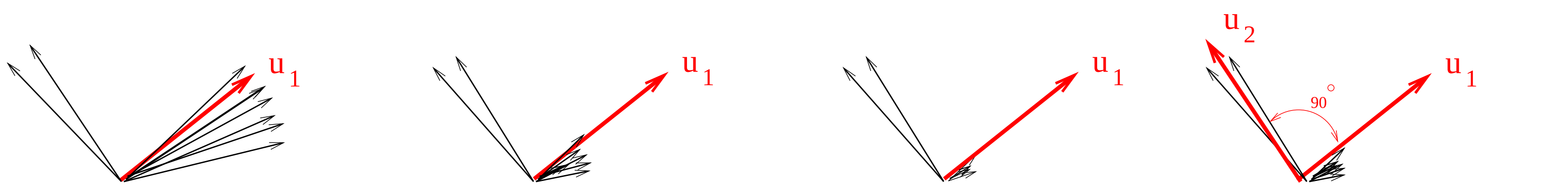
Repeat to get next u (orthogonal to previous u_i's)

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 qth 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: $\sum_t (\text{Dom}(t)^2/n)^2$

- one giant topic $\rightarrow 1$
- k same-size topics with no document sharing $\rightarrow 1/k$

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)