24. Dimensionality Reduction & Latent Semantic Analysis

INFO 202 - 19 November 2008

Bob Glushko

Plan for Today's Class

Limitations of the Vector Model

Linguistic vs Statistical Approaches in Natural Language Processing

Dimensionality Reduction -- Intuitive Motivation and Description

Dimensionality Reduction -- Technical Description

Latent Semantic Analysis and Applications

Reminder: Models of Information Retrieval

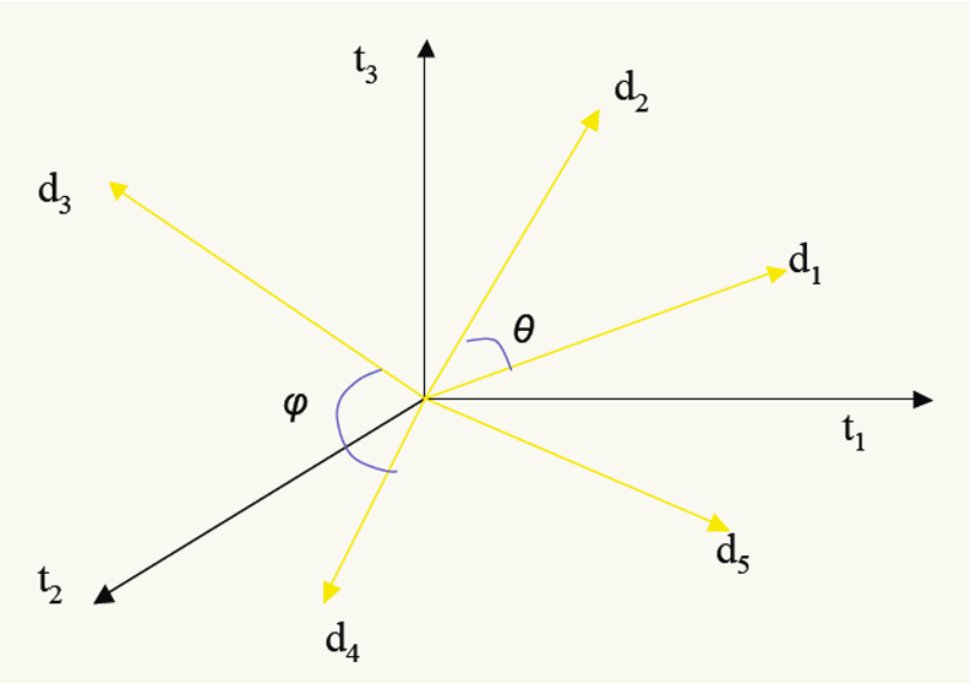
The core problems of information retrieval are finding relevant documents and ordering the found documents according to relevance

The IR model explains how these problems are solved:

- ...By specifying the representations of queries and documents in the collection being searched
- ...And the information used, and the calculations performed, that order the retrieved documents by relevance
- (And optionally, the model provides mechanisms for using relevance feedback to improve precision and results ordering)

Different IR models solve these problems in different ways; there is usually a tradeoff that the better they solve them, the more computationally complex they are

The Vector Model



Vector Model: Advantages

Index terms can be selected automatically

Term weighting to improve retrieval performance

Partial matching of queries and documents when no document contains all search terms

Relevance ranking according to similarity

Relevance feedback incorporated by modifying query vector

Vector Model: Limitations

The calculations used by simple vector models are about the frequency of words and word forms (e.g., stemmed) in texts

This means that they are measuring the "surface" usage of words as patterns of letters

They can't distinguish different meanings of the same word (polysymy)

They can't detect equivalent meaning expressed with different words (synonymy)

Polysymy in the Vector Model

Because the vector model doesn't recognize that "BANK as in river" and "BANK as in money" are different senses, all occurrences of the term BANK are treated the same instead of being distinguished as separate dimensions in the space

This overestimates the similarity of documents containing BANK

$$\mathrm{sim}_{\mathrm{true}}(d, q) < \cos(\angle(\vec{d}, \vec{q}))$$

Synonyms in the Vector Model

The vector model can't recognize that "AUTO" and "CAR" are synonyms, and thus assigns them separate dimensions instead of counting them as additional occurrences of the same "semantic term"

This underestimates the similarity of documents containing AUTO and CAR

$$\mathrm{sim}_{\mathrm{true}}(d, q) > \cos(\angle(\vec{d}, \vec{q}))$$

Flashback to September 29...

I saw a:	with a:
Man	Telescope
Star	Microscope
Molecule	Binoculars

How many combinations make sense?

Language and Meaning

Words and sentence structure only hint at meaning

Meaning is constructed from all the clues or cues in the context of use -- common knowledge, assumptions, previous discourse, the present situation, and inferences from all of these

How much "context" and "common knowledge" must be represented / understood to make sense of what meaning is intended?

A great deal of work in artificial intelligence has been dedicated to building knowledge bases to support language understanding, reasoning, problem solving applications

Do we need to use this kind of knowledge to solve the polysymy and synonymy challenges in information retrieval?

Two Radically Different Approaches to Natural Language Processing

LINGUISTIC Approach:

Linguistic models of grammar, morphology, and phonology are essential prerequisites for NLP

We also need to develop models of the "human language processor" and combine them with the linguistic models

STATISTICAL Approach:

Statistical analysis of language reveals structure and patterns

This extracted knowledge -- represented as the occurrence or co-occurrence probabilities of specific things -- can answer many of the questions that supposedly require "understanding" or more abstract "rules"

Motivating Data-Driven Language Study

Early Data-driven Perspectives on Language [1]

Just as in the last 10-15 years, as cognitive science has emerged in the intersection of cognitive psychology, linguistics, and computer science, if we look back to when computers were first being invented around 1940 the study of language reflected the prevailing psychology theories of behaviorism and structuralism

These theories held that language was learned through empirical learning mechanisms of conditioning, association, practice in exercising skills

These stimulus -> response notions do not postulate any internal knowledge representation. This perspective on language suggests a statistical approach to NLP

Early Data-driven Perspectives on Language [2]

GK Zipf first expressed "Zipf's Law" in a 1935 book titled *Psycho-Biology of Language*

All of the work in WWII on codebreaking and cryptography emphasizes the empirical study of word and language patterns (Turing)

Claude Shannon on information theory (1948) says that the information in a message is not defined by its content but by its probability of being chosen among several alternatives

Chomsky's Argument for "Deep" Language Analysis [1]

In late 1950s the statistical approaches fell out of favor, mostly because of the work of Noam Chomsky (*Syntactic Structures*, 1957)

- *Colorless green ideas sleep furiously*
- *Furiously sleep ideas green colorless*

Neither of these sentences is natural language and won't occur in language samples, but the former is grammatical and the second isn't

Chomsky's Argument for "Deep" Language Analysis [2]

So language knowledge can't be based on learned behavior -- because the relevant data is sparse (many reasonable sentences never appear); instead, it is generative and based on rules and representations

This view of language as a formal, mathematically-describable system became the dominant view in linguistics and computer science

Probabilistic Models: The Data Strikes Back

These spoken phrases can be acoustically identical:

- *Your lie cured mother*
- *You like your mother*

Starting in the late 1980s, speech recognition systems used huge sets of speech data to build probabilistic models

NLP by Computers != NLP by People

The "rules of language" are a theory of the knowledge that fluent speakers possess (competence), not a theory of how they generate and understand language (performance)

We have no conscious awareness of how we process language, and while we can sometimes explicitly apply the "rules" it certainly doesn't seem that we use that kind of abstract information in "normal" NLP

But likewise, it doesn't seem that we explicitly use information about the likelihood of various language structures occurring or co-occurring, which is what statistical NLP does

"Learning the statistics" != "Statistically-driven learning"

What people seem to do as they learn language is "statistically-driven learning," not "learning the statistics"

That is, they use statistical evidence to build knowledge about the language into internal representations and language processing mechanisms that are more general than the specific data they were built with

Human language processing has been successfully modeled using neural nets and other representations in which the "statistics" are encoded in the pattern of activations in a distributed way

Combined Approaches

Today both the linguistic and data-driven approaches are seen as integral and complementary parts of an NLP application

Systems employ sophisticated techniques for dictionaries and grammars to identify parts of speech and do morphological analysis

But the statistics of co-occurrence / conditional probability yield many practical techniques for estimating the substitutability or semantic equivalence of words in larger text segments that make no use of their "languageness"

In particular, the web is such a huge corpus that statistical approaches can be surprisingly informative and robust

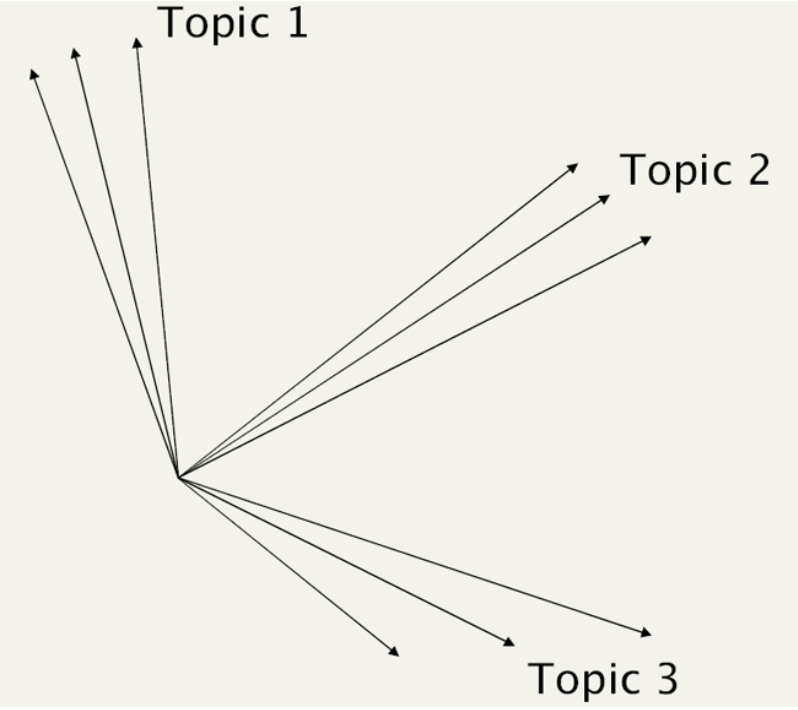
Dimensionality Reduction to Improve the Vector Model

The dimensionality of the space in the simple vector model is the number of different terms in it

But the "semantic dimensionality" of the space is number of distinct topics represented in it

The number of topics is much lower than the number of terms (in a given collection, untapped synonymy is more important than unnoticed polysymy)

"Topic Space," Not "Term Space"



Example: Word Sense Disambiguation using Lexical Co-Occurrences

The co-occurrences of words in a text collection can tell us what the documents are about and distinguish different senses of polysemous words

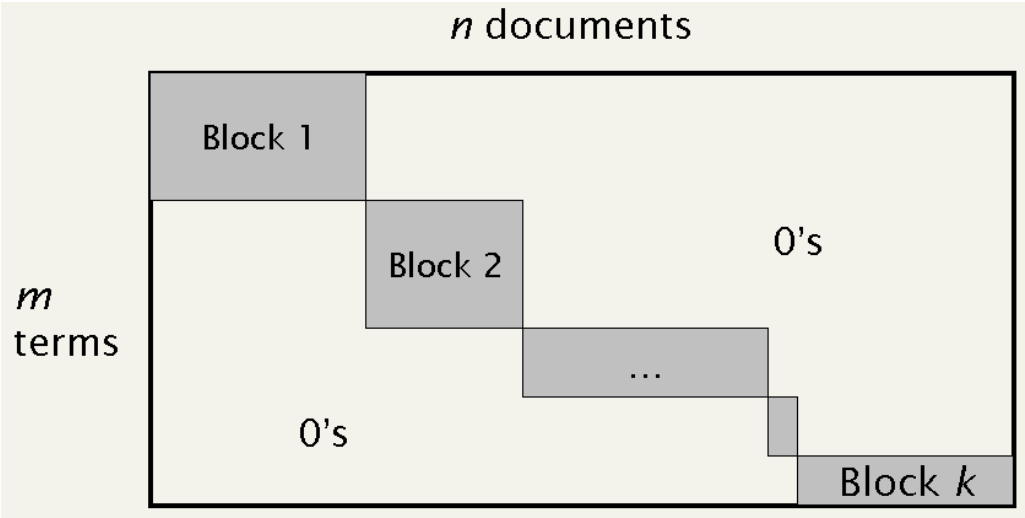
Co-occurrences of frequent words are uninteresting:

- "doctor" co-occurs with "with," "a," and "is"

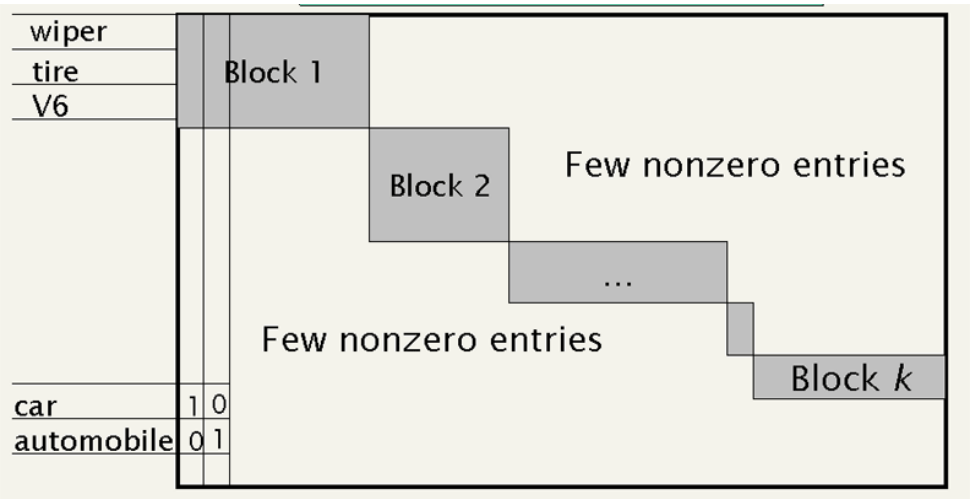
Co-occurrences between two less frequent words help up understand what a text is about:

- "doctor" co-occurs with "honorary," "dentist," "nurse," "examine, "treat," etc.
- "drug" co-occurs with "price," "prescription" and "patient" in medical contexts and with "abuse," "paraphernalia," and "illicit" in non-medical contexts

An Intuitive Explanation for Dimensionality Reduction Techniques

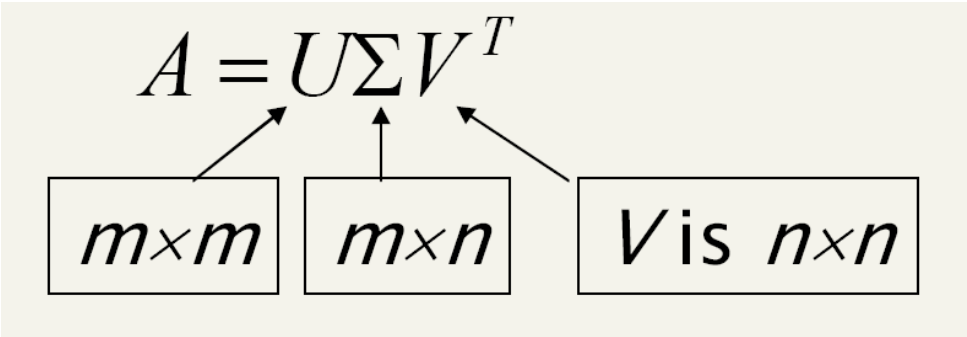


An Intuitive Explanation for Dimensionality Reduction Techniques



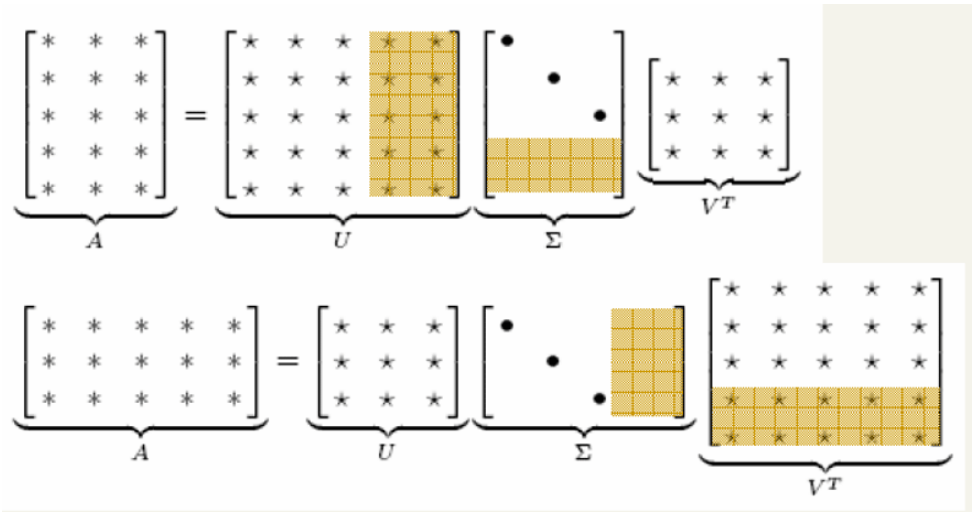
Singular Value Decomposition

For an m× n matrix A of rank r there exists a factorization (Singular Value Decomposition = SVD) as follows:



The original matrix is decomposed here into three matrices - two that factor the original rows and columns into orthogonal vectors, and a diagonal matrix that contains scaling values

Illustrating SVD and Sparseness



Dimensionality Reduction with "Latent Semantic Analysis"

Once we have factored the original term x document matrix using SVD, we can then find a much smaller matrix that approximates it

(...more or less by deleting coefficients from the diagonal matrix, starting with the smallest)

These techniques in effect "squeeze down" the matrix to lower rank (typically 100-300) by bringing together terms that have similar co-occurrence patterns

The vectors in this reduced dimensionality space aren't directly identifiable as any lexical or semantic component, but they are "latently" semantic in that relationships between vectors in this lower dimensional space reflect semantic associations

LSA as an IR Model

Reducing the dimensionality of the term x document matrix means we are discarding some of the descriptors applied to each document in the collection, which might suggest that retrieval precision would suffer

But we're not just discarding terms -- we are replacing sets of co-occurring (e.g., associated) terms with "superterms" or "topics" that represent meaning as a kind of average of all the terms that tend to occur in the same contexts

So we can compute document similarity based on the inner product / cosines in this latent semantic space

Two Views of LSA

LSA has been shown to be a practical technique for estimating the substitutability or semantic equivalence of words in larger text segments

In addition, some of its proponents (e.g., Dumais) view it as a model of the computational processes and representations underlying substantial portions of how knowledge is acquired and used

And while it is highly unlikely that the human brain uses the same mathematical algorithms as LSA/SVD, it is almost certain that the brain uses as much analytic power to transform temporally localized experiences into synthesized knowledge

LSA Demonstrations (lsa.colorado.edu)

"Nearest Neighbors" -- find words that are "near" a text sample in LSA space (even if they don't appear in the sample)

"One to Many" Comparisons -- evaluate the similarity of texts to a given text

Nearest Neighbors Example 1

INPUT: (From chapter 1 of Glushko & McGrath "Document Engineering"): In the 19th century the telegraph and telephone made it possible to exchange information electronically and coordinate business activities at a scale vastly larger than before, leading to the rise of the modern corporation. The late 20th and early 21st centuries have witnessed the equally profound impact of the Internet (and related technologies such as the World Wide Web, electronic mail, and XML) on how businesses work. Now the web-based virtual enterprise can be open for business 24 hours a day, 7 days a week, with a global presence enabled by distributing people and resources wherever they are needed in either physical space or cyberspace.

NEIGHBORS (in ranked order): business, information, telecommunications, dun, database, videotext, videotex, telephone, teletext, bradstreet, activities, retrieval, ibm, bookkeeping, technologies, accounting, microelectronics, communications, advertising, consumers

Nearest Neighbors Example 2

INPUT (From Saxenian "): Traditional theories of economic development assume that new products and technologies emerge in industrialized nations that can combine sophisticated skills and research capabilities with large, high-income markets, and that mass production is shifted to less costly locations once the product is standardized and the manufacturing process has matured. In this view, success in the periphery builds on the success of more advanced economies: late developers are destined to remain followers because leading-edge skills and technology reside in the corporate research labs and universities in the core.

NEIGHBORS (in ranked order): development, research, production, consumers, demand, technology, product, economic, technologies, economics, stagflation, market, process, capitalism

One to Many Comparison

The Target Text: Sample answer from question 1 on 2007 202 midterm (Memex v. Del.icio.us)

Comparison Texts: 2 student answers, Glushko & McGrath text, Saxenian text

RESULTS (similarity to Text 1): Text 2 0.96, Text 3 0.93, Text 4 0.89, Text 5 0.79

LSI's Limitations

The computational cost of creating the lower dimensional matrix is significant

No collection has had more than 1 million documents (this may seem like a lot, but is tiny compared to the Web)

And like simple vector models, there is no good way to express negations in queries

Reading for Lecture #25 on 24 November

Alejandro Diaz, "Through the Google Goggles: Sociopolitical Bias in Search Engine Design." 2005 (Chapters 5 & 6)

Manning et al., Chapter 21

Pairin Katerattanakul, Bernard Han, and Soongoo Hong. "Objective Quality Ranking of Computing Journals," 2003