UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

# INFO 202
# "Information Organization & Retrieval"
# Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

12 December 2013
Lecture 30.1 – Course Review

# IR Models

- 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)

# IR Models

- BOOLEAN model -- representations are sets of index terms, set theory operations with Boolean algebra calculate relevance as binary
- VECTOR models -- representations are vectors with non-binary weighted index terms, linear algebra operations yield continuous measure of relevance

# IR Models

- STRUCTURE models -- combine representations of terms with information about structures within documents (i.e., hierarchical organization) and between documents (i.e. hypertext links and other explicit relationships) to determine which parts of documents and which documents are most important and relevant
- PROBABILISTIC models -- documents are represented by index terms, and the key assumption is that the terms are distributed differently in relevant and non relevant documents.

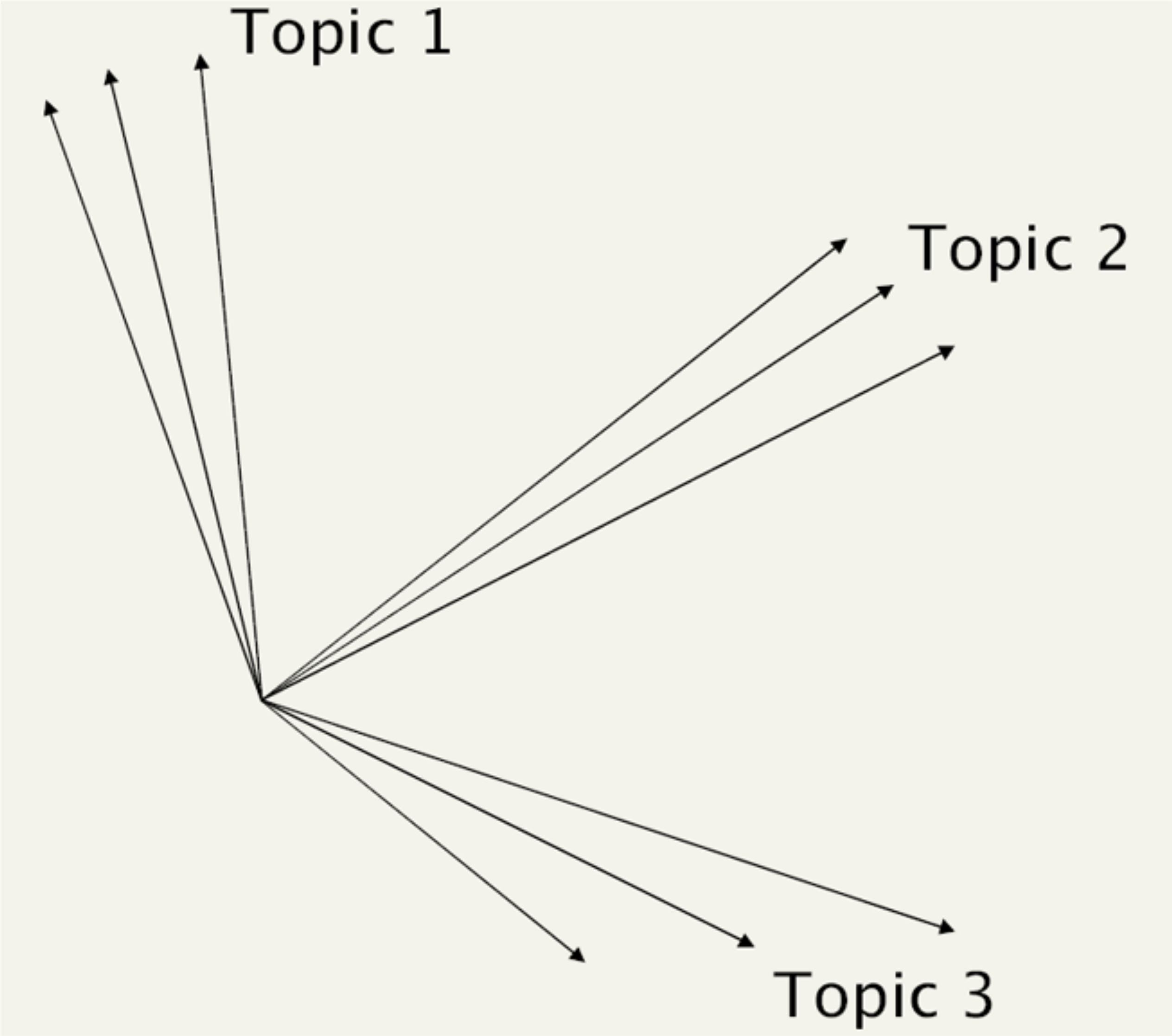
# Dimensionality Reduction - A Very Informal Motivation

- If every resource described as “big” is also described as “red,” and every “small” resource is also “green,” this correlation between color and size means that either of these properties is sufficient
- With thousands of properties or descriptive terms, we need clever statistical analysis to choose the optimal descriptive terms
- We can synthesize  new “logical terms” based on the correlations

# From Terms to Topics

- 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 the number of distinct topics represented in it
- The number of topics is much lower than the number of terms
- Documents can be similar in the topics they contain even if they have no words in common

# "Topic Space," Not "Term Space"



# Relevance in the Boolean Model

- Because terms are either present or absent, a document is either relevant or not relevant
- Retrieved documents might be ordered (chronologically?) but not by relevance because there is logically no way to calculate it
- But this is clearly flawed -- if a query is "a and b" the Boolean model retrieves only those documents that contain both of them, and treats documents that contain only a or only b as equally irrelevant

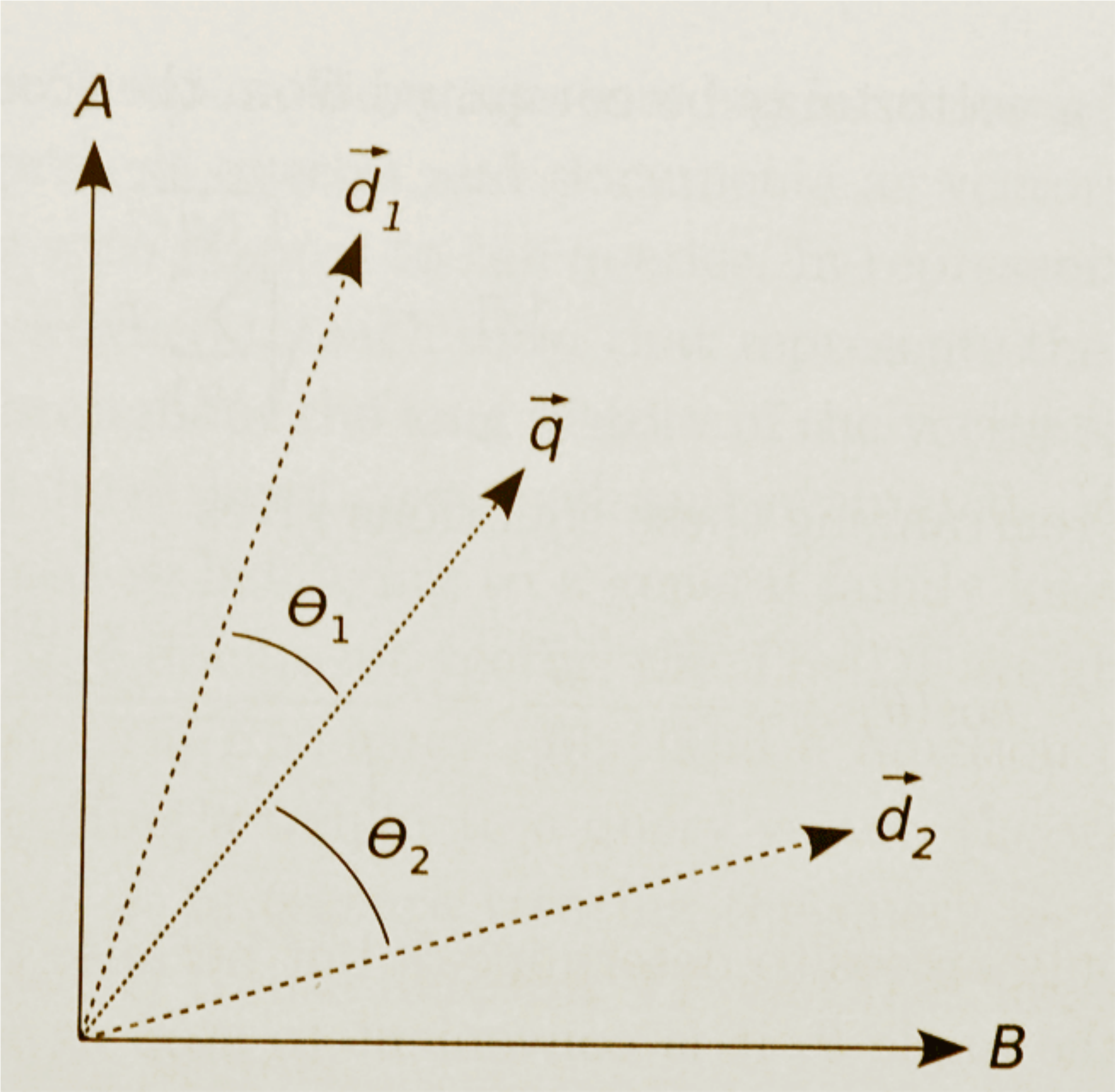
# Relevance in other IR Models

- Other IR models rank the retrieved documents in terms of their relevance to the query
- This requires more refined methods of representing what documents are about
- This enables more continuous methods of assessing relevance rather than all or none

# Vector Model Retrieval and Ranking

- So vector queries are fundamentally a form of "evidence accumulation" where the presence of more query terms in a document adds to its "score"
- This score is not an exact measure of relevance with respect to the query, but it is vastly better than the all or none Boolean model!

# Similarity in Vector Models (Graphical Depiction)



# IR Models and “Information Needs”

- In classical IR, an "information need" is formulated as a query submitted to some information collection
- "Documents" in the information collection are returned that "satisfy" the query
- The "IR model" specifies how the queries and documents are represented and how "satisfaction" is calculated
- "Satisfaction" can be an exact match, but more often the returned documents are ranked according to the statistical similarity between their representations and that of the query

# Getting Beyond the “10 Blue Links”

- IR systems locate relevant documents in response to a query, but the user must extract the actual answer to his or her question
- This challenge has been described as “getting beyond the 10 blue links” – here are some of the ways to do that:
    - Semantic search
    - Structured data search
    - Structure  / graph search

# Crossing the Semantic Gap Through Computation

- A consequence of the semantic gap for mulitimedia is that there are a very large number of low-level features that can be reliably identified
- Any description using these features will be "sparse" - lots of missing values
- Dimensionality reduction techniques can exploit correlations between low-level features

# Natural Language Processing

- Every NLP application needs to identify and classify the words, phrases, structures, documents  in some context or domain
- The earliest NLP applications were very narrow in scope, relying on "hand crafted" rules to implement the required knowledge
- Today much NLP uses dictionaries, stemmers, grammars, and other language knowledge
- But the statistics of co-occurrence / conditional probability yield many practical techniques for analyzing and generating language that make no use of "languageness"

# Real World NLP Applications

- Information extraction
- Summarization
- Auto-journalism (story generation)
- Machine translation
- Speech recognition and synthesis
- Computational classification
    - Topic identification
    - Author identification
    - Spam detection
    - Sentiment analysis
- Question answering, automated customer service, recommendations…

# Computational Classification

- Some classification tasks can't be done by people at the needed scale or speed
- Some of the tasks can be done by computational approaches like N-gram analysis that are very simple yet very useful
- Some tasks are inherently difficult and require techniques of “machine learning”, which are not simple

# Text Classification Problems

- Classification assigns objects in some domain to one of at least two classes or categories
  - words - determine part of speech
  - words - disambiguate polysemy
  - document retrieval - relevant/not relevant?
  - author identification - Shakespeare or not?
  - sentiment classification - positive or negative affect? urgent or not urgent?
  - language - English, Spanish, whatever?

# Computational Classification

- CLASSIFICATION assumes a system of categories and some labeled instances so we can train a system to assign new instances to the appropriate categories
- In contrast, CLUSTERING techniques don't assume pre-existing categories - they create them (usually to maximize similarity within categories and maximize it between them)

# Understanding Information Extraction

- What type of structure is being extracted?
- What is the unstructured source input?
- What resources are available to guide the extraction?
- What extraction techniques are employed?
- What is the format of the extracted information?

(IE typically starts after tokenization, part-of-speech tagging, and phrase identification steps of text processing)

# Sentiment Analysis

- Sentiment analysis (aka "opinion mining") can be thought of a three-stage classification problem
  - Entity extraction to locate text of interest
  - Classifying texts as opinions or facts
  - Classifying the opinions according to polarity - positive vs. negative (or on some numerical scale)
- This is challenging because these classes are really continuaa without sharp boundaries
- ...and because sarcasm, slang, cliches, and cultural norms obscure the content used to make the classification

# The Organizing System Lifecycle

- There is always a lifecycle, but there are times when its phases need to be more explicit and formal:
  - In institutional contexts
  - In information-intensive contexts
  - When traceability and impact analysis are necessary
- Better to be more explicit and formal than absolutely necessary than vice versa

# Defining and Scoping the Domain

- Determining scope and scale
- Nature and number of users
- Expected lifetime
- Physical and technological environment
- Relationship to other organizing systems

# Physical or Technological Environment

- There might be affordances that create possibilities
- But there might be constraints that limit them
- Estimating the ultimate size of a collection at the beginning of an organizing system's lifecycle can reduce scaling issues related to storage space for the resources or for their descriptions (flashback to "warrant" goes here)

# Requirements for Interactions

- All organizing systems have some common interactions, but most of the time we want to pay attention to the more resource-specific interactions that create the most value
- The priorities of different interactions are often determined by decisions about intended users
- An essential requirement is ensuring that the supported interactions can be discovered and invoked by their intended users

# Requirements for Interactions

- For most organizing systems other than personal ones, the set of interactions that are implemented in an organizing system is strongly determined by business model considerations, funding levels, or other economic factors

- Businesses differentiate themselves by the number and quality of the interactions they support with their resources

# Requirements about Resource Description

- The most generic interactions use descriptions that can be associated with almost any type of resource, such as the name, creator, and date
- Different types of resources must have differentiating properties, otherwise there would be no reason to distinguish them

# Requirements about Resource Description

- Business strategy and economics strongly influence the extent of resource description and the use of technology for automatic description
- The tradeoffs imposed by the extent and timing of resource description arise throughout the lifecycle, with the tradeoff between recall and precision being the most salient

# Tradeoffs involving Description

- Someone designs or selects a structure for the description... or not
- Someone determines the content of the description .. or not
- How much structure or how detailed a description?

# Tradeoffs involving Description

- Is it easier to create structured or unstructured descriptions?
- If we want to combine information from many different authors or sources, what are the implications for description and organizing decisions?
- Is it easier to combine information from different authors or sources if it is structured or unstructured?

# The Fundamental Tradeoff in an Organizing System

- There is a tradeoff between the amount of work that goes into describing and organizing a collection of resources and the amount of work required to find and use them
- The more effort we put into describing and organizing resources, the more effectively they can support interactions
- The more effort we put into retrieving resources, the less they need to be organized first

# The Fundamental Tradeoff in an Organizing System

- We need to think in terms of investment, allocation of costs and benefits between the describer/organizer and user
- The allocation differs according to the relationship between them; who does the work and who gets the benefit?

# The Course In One Slide

- To organize is to create capabilities by intentionally imposing order and structure
- We organize things, we organize information, we organize information about things, and we organize information about information
- If we think abstractly about these activities, we can see commonalities that outweigh their differences; We select, organize, interact with, and maintain resources
- We organize resources as individuals, in informal association with other individuals, or as part of a more formal institutional or business context
- We must recognize the profound impact of new technologies and their co-evolution with the nature of the organizing we do and the kinds of interactions that this organizing enables, but can't ignore the "classical" concepts and knowledge

# Your Final Exam

- Tuesday December 17, 9-1 – Early Final
- Wednesday December 18, 9-1 – FINAL EXAM

# My Final Exam

- On the first day of class I said:
  - We deal with deep intellectual issues that have challenged philosophers and other deep thinkers for millennia
  - You must make the transition to studying information / content IN a discipline to studying information / content AS a discipline
  - You must learn to look past the presentation / rendition / technology reification / thinginess of information to see it more abstractly as structure and meaning

# My Final Exam

- On the first day of class I said:
  - This course WILL change how you think about information
  - If it doesn't, it means that we have both failed this semester

# This is NOT the end of 202

- It IS the end of the semester: 1500 pages, 10 assignments, 28 lectures, 12 section meetings…
- But how you think and talk about information and organization has changed immensely
- This transformation will shape the remainder of your ISchool experience and the rest of your personal and professional lives
- Thank you for giving me the privilege of being along for the ride
- Good luck on the final exam