Search Engines

Information Retrieval in Practice

Beyond Bag of Words

- "Bag of Words"
  - a document is considered to be an unordered collection of words with no relationships
- Extending representation
  - feature-based models
  - dependency models
  - document structure
  - question structure
  - other media

Feature-Based Retrieval Models

- *Linear* feature-based model

$$S_\Lambda(D;Q) = \sum_j \lambda_j \cdot f_j(D,Q) + Z$$

$S_\Lambda(D;Q)$ is a scoring function for a document D and query Q, parameterized by Λ
$f_j(D,Q)$ is a feature function that maps query/document pairs to real values
Z is a constant that does not depend on D (but may depend on Λ or Q)

- Some models support non-linear functions, but linear is more common

Linear Feature-Based Models

- To find best values for parameters
  - need a set of training data *T*
  - an evaluation function $E(\mathcal{R}_\Lambda; \mathcal{T})$
    - where R_Λ is the set of rankings produced by the scoring function for all the queries
- Goal of a linear feature-based retrieval model is to find a parameter setting that maximizes *E* for the training data

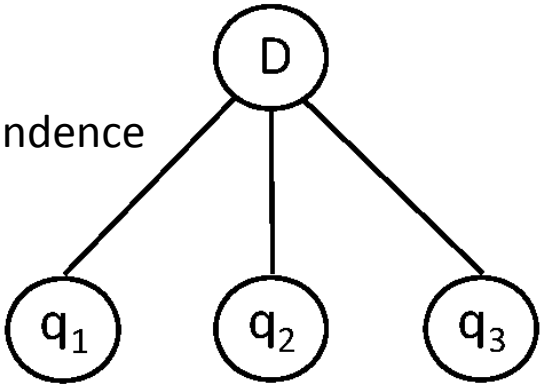
$$\widehat{\Lambda} = \arg\max_\Lambda E(\mathcal{R}_\Lambda; \mathcal{T})$$

Term Dependence Models

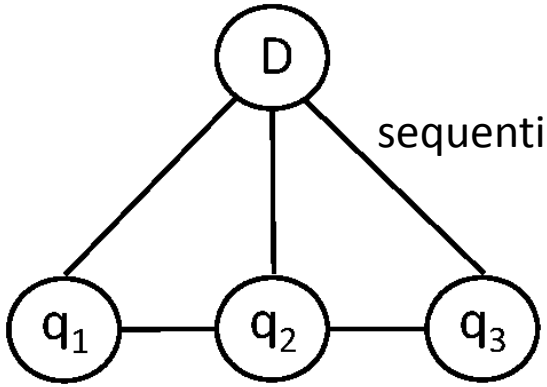
- Term dependence models do not assume that words occur independently of each other
- e.g., *Markov Random Field (MRF)* model
  - construct a graph that consists of a document node and one node per query term
  - *undirected graphical model*
  - models the joint distribution over the document random variable and query term random variables
  - models dependencies between random variables by drawing an edge between them

MRF Model Assumptions

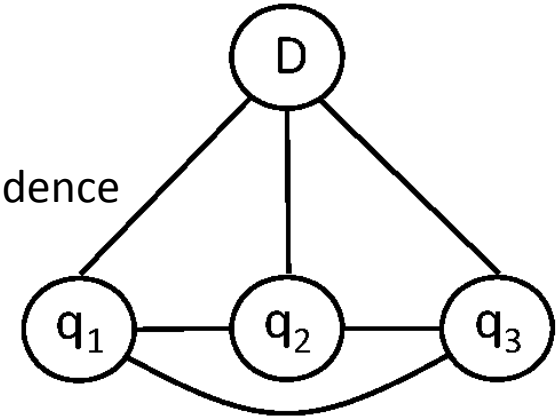
independence



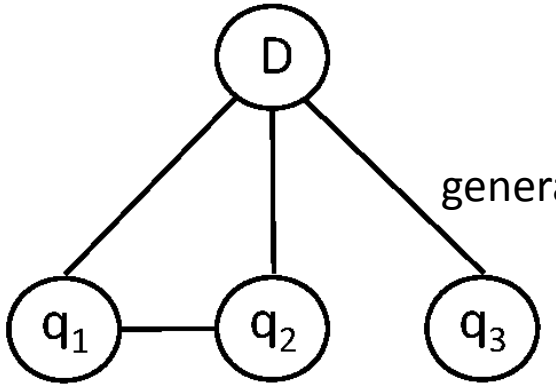
sequential dependence



full dependence



general dependence



MRF Model

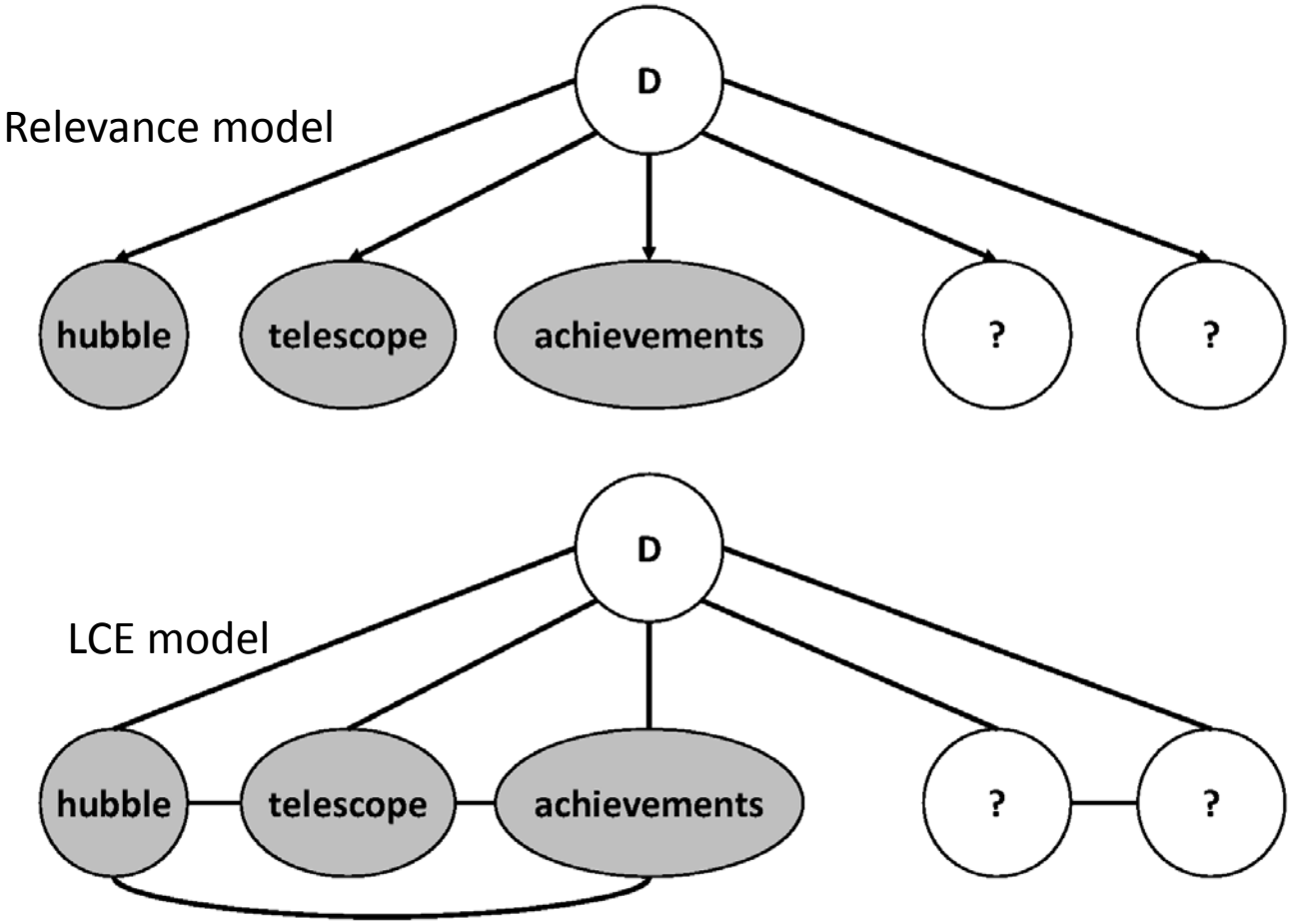
- Define a set of *potential functions* over the *cliques* of the graph
  - these are the features of the linear feature-based model
  - e.g. sequential dependence model in Galago for query “president abraham lincoln”

```
#weight(0.8 #combine(president abraham lincoln)
            0.1 #combine(#od:1(president abraham)
                         #od:1(abraham lincoln)
            0.1 #combine(#uw:8(president abraham)
                         #uw:8(abraham lincoln)
```

Latent Concept Expansion

- Generalized version of pseudo-relevance feedback and relevance models
  - Latent, or hidden, concepts are words or phrases that users have in mind but do not mention explicitly when they express a query
  - latent concept expansion graph shows dependencies between query words and expansion words
  - better probability estimates for expansion terms
  - expansion *features* not just terms

Pseudo-Relevance Feedback Graphs



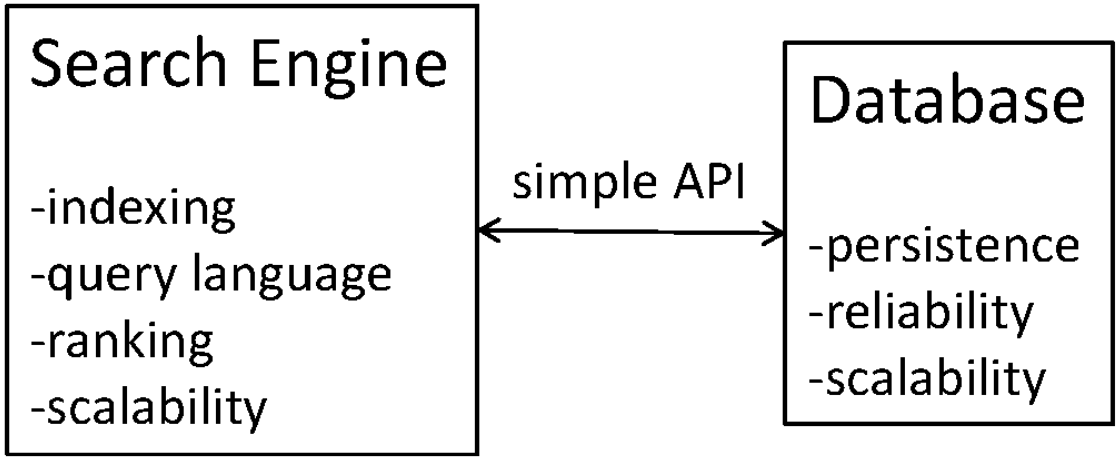
LCE Example

1-word concepts	*2-word concepts*
telescope	hubble telescope
hubble	space telescope
space	hubble space
mirror	telescope mirror
NASA	telescope hubble
launch	mirror telescope
astronomy	telescope NASA
shuttle	telescope space
test	hubble mirror
new	NASA hubble
discovery	telescope astronomy
time	telescope optical
universe	hubble optical
optical	telescope discovery
light	telescope shuttle

Integrating Databases and IR

- Possible approaches
  - Extending a database model to more effectively deal with probabilities
  - Extending an information retrieval model to handle more complex structures and multiple relations
  - Developing a unified model and system
- Applications such as web search, e-commerce, and data mining provide testbeds

Interaction of Search and Databases



e.g., e-commerce applications such as Amazon

XML Retrieval

- XML is an important standard for both exchanging data between applications and encoding documents
- Database community has defined languages for describing the structure of XML data (*XML Schema*), and querying and manipulating that data (*XQuery* and *XPath*)
  - query languages similar to SQL but must handle hierarchical structure
  - XPath restricted to single document type

XML Retrieval

- *INEX* project studies XML retrieval models and techniques
  - similar evaluation approach to TREC
  - queries are specified using a simplified version of XPath called *NEXI*
  - NEXI constructs include *paths* and *path filters*
    - A path is a specification of an element (or node) in the XML tree structure
    - A path filter restricts the results to those that satisfy textual or numerical constraints

NEXI Examples

//A//B - any B element that is a descendant of an A element in the XML tree. A descendant element will be contained in the parent element.

//A/* - any descendant element of an A element.

//A[about(.//B,"topic")] - A elements that contain a B element that is about "topic". The about predicate is not defined, but is implemented using some type of retrieval model. .//B is a *relative* path.

//A[.//B = 777] - A elements that contain a B element with value equal to 777.

INEX Examples

//article[.//fm/yr < 2000]//sec[about(.,"search engines")]
- find articles published before 2000 (fm is the front matter of the article) that contain sections discussing the topic "search engines".

//article[about(.//st,+comparison) AND about (.//bib,"machine learning")]
- find articles with a section title containing the word "comparison" and with a bibliography that mentions "machine learning".

//*[about(.//fgc, corba architecture) AND about(.//p, figure corba architecture)]
- find any elements that contain a figure caption about "corba architecure" and a paragraph mentioning "figure corba architecture".

Entity Search

- Identify entities in text
- Construct “pseudo-documents” to represent entities
  - based on words occurring near the entity over the whole corpus
  - also called “context vectors”
- Retrieve ranked lists of entities instead of documents

Entity Search Example

Query:

biomedical research and technology

Top Ranked Results:

minneapolis research
signs inc.
syntex
california institute of technology
massachusetts institute of technology
therapeutic products

(organization search based on a TREC news corpus)

Expert Search

- Find "experts" for a given topic
  - recent TREC track
- Rank candidate entities *e* by the joint distribution *P(e, q)* of entities and query terms

$$P(e,q) = \sum_{d \in D} P(e,q|d)P(d)$$

$$P(e,q|d) = P(q|e,d)P(e|d)$$

  - *P*(*q*|*e,d*) involves ranking entities in those documents with respect to a query
  - *P*(*e*|*d*) component corresponds to finding documents that provide information about an entity

Expert Search

- Assuming words and entities are independent leads to poor performance

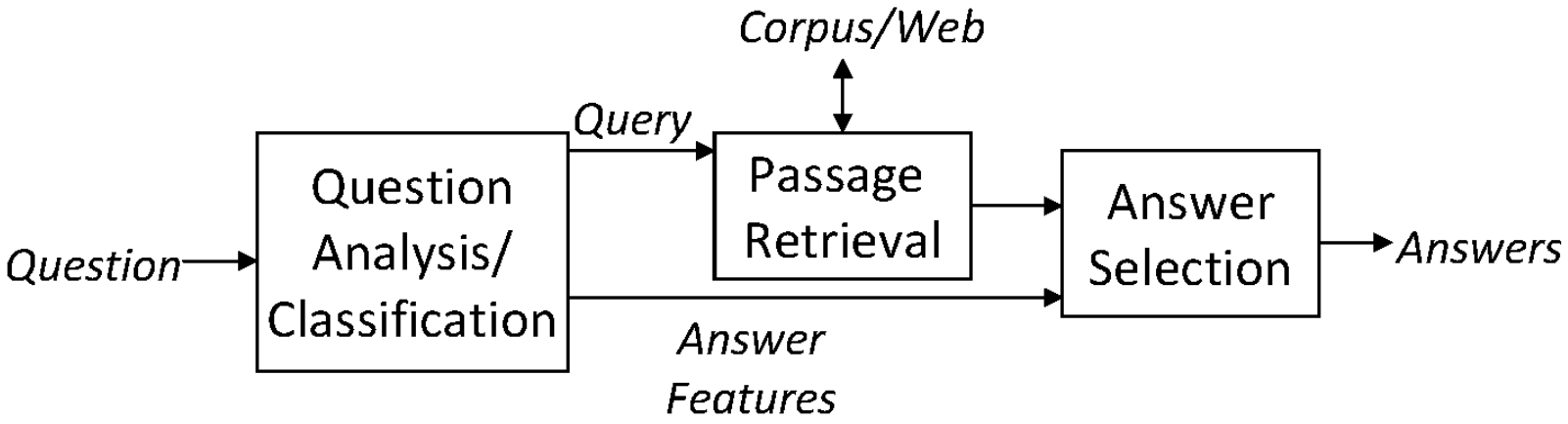
$$P(e, q|d) = P(q|d)P(e|d)$$

- Instead estimate the strength of association between *e* and *q* using proximity of co-occurrence of the query words and the entities

Question Answering

- Providing *answers* instead of ranked lists of documents
- Older QA systems *generated* answers
- Current QA systems *extract* answers from large corpora such as the Web
- *Fact-based* QA limits range of questions to those with simple, short answers
  - e.g., *who*, *where*, *when* questions

QA Architecture



Fact-Based QA

- Questions are *classified* by type of answer expected
  - most categories correspond to named entities
- Category is used to identify potential answer passages
- Additional natural language processing and semantic inference used to rank passages and identify answer

Example Question	*Question Category*
What do you call a group of geese?	Animal
Who was Monet?	Biography
How many types of lemurs are there?	Cardinal
What is the effect of acid rain?	Cause/Effect
What is the street address of the White House?	Contact Info
Boxing Day is celebrated on what day?	Date
What is sake?	Definition
What is another name for nearsightedness?	Disease
What was the famous battle in 1836 between Texas and Mexico?	Event
What is the tallest building in Japan?	Facility
What type of bridge is the Golden Gate Bridge?	Facility Description
What is the most popular sport in Japan?	Game
What is the capital of Sri Lanka?	Geo-Political Entity
Name a Gaelic language.	Language
What is the world's highest peak?	Location
How much money does the Sultan of Brunei have?	Money
Jackson Pollock is of what nationality?	Nationality
Who manufactures Magic Chef appliances?	Organization
What kind of sports team is the Buffalo Sabres?	Org. Description
What color is yak milk?	Other
How much of an apple is water?	Percent
Who was the first Russian astronaut to walk in space?	Person
What is Australia's national flower?	Plant
What is the most heavily caffeinated soft drink?	Product
What does the Peugeot company manufacture?	Product Description
How far away is the moon?	Quantity
Why can't ostriches fly?	Reason
What metal has the highest melting point?	Substance
What time of day did Emperor Hirohito die?	Time
What does your spleen do?	Use
What is the best-selling book of all time?	Work of Art

Other Media

- Many other types of information are important for search applications
  - e.g., scanned documents, speech, music, images, video
- Typically there is no associated text
  - although *user tagging* is important in some applications
- Retrieval algorithms can be specified based on any content-related features that can be extracted

Noisy Text

- OCR and speech recognition produce *noisy* text
  - i.e., text with numerous errors relative to the original printed text or speech transcript
- With good retrieval model, effectiveness of search is not significantly affected by noise
  - due to *redundancy* of text
  - problems with short texts

OCR Examples

Original:

The fishing supplier had many items in stock, including a large variety of tropical fish and aquariums of all sizes.

OCR:

The fishing supplier had many items in stock, including a large variety of tropical fish and aquariums ot aH sizes~

Original:

*** This work was carried out under the sponsorship of National Science Foundation Grants NSF-GN-380 (Studies in Indexing Depth and Retrieval Effectiveness) and NSF-GN-432 (Requirements Study for Future Catalogs).**

OCR:

This work was carried out under the sp011J!0rship 01 NatiolUl1 Setenee Foundation 0rant. NSF·0N·SB0 (Studl .. In Indexing Depth and Retrieval Eflccth"ene&&) and NSF·0N·482 (Requirements Study lor Future 'Catalogs)•

Speech Example

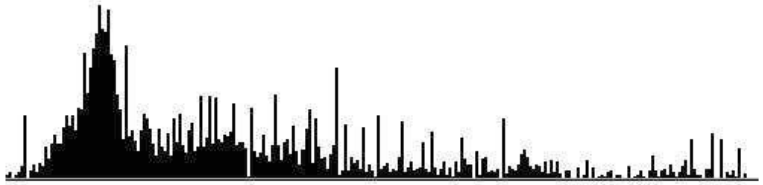
Transcript:
French prosecutors are investigating former Chilean strongman Augusto Pinochet. The French justice minister may seek his extradition from Britain. Three French families whose relatives disappeared in Chile have filed a Complaint charging Pinochet with crimes against humanity. The national court in Spain has ruled crimes committed by the Pinochet regime fall under Spanish jurisdiction.

Speech recognizer output:
french prosecutors are investigating former chilean strongman of coastal fish today the french justice minister may seek his extradition from britain three french families whose relatives disappeared until i have filed a complaint charging tenants say with crimes against humanity the national court in spain has ruled crimes committed by the tennessee with james all under spanish jurisdiction

Images and Video

- Feature extraction more difficult
- Features are low-level and not as clearly associated with the semantics of the image as a text description
- Typical features are related to color, texture, and shape
  - e.g., color histogram
    - “quantize” color values to define “bins” in a histogram
    - for each pixel in the image, the bin corresponding to the color value for that pixel is incremented by one
  - images can be ranked relative to a query image

Color Histogram Example



peak in yellow

Texture and Shape

- Texture is spatial arrangement of gray levels in the image
- Shape features describe the form of object boundaries and edges
- Examples:

← shape

Video

- Video is segmented into shots or scenes
  - continuous sequence of visually coherent frames
  - boundaries detected by visual discontinuities
- Video represented by *key frame* images
  - e.g., first frame in a shot

Image Annotation

- Given training data, can learn a joint probability model for words and image features
- Enables automatic text annotation of images
  - current techniques are moderately effective

people, pool, swimmers, water

cars, formula, tracks, wall

clouds, jet, plane, sky

fox, forest, river, water

errors

Music

- Music is even less associated with words than images
- Many different representations
  - e.g., audio, MIDI, score
- Search based on features such as spectrogram peaks, note sequences, relative pitch, etc.