CS347

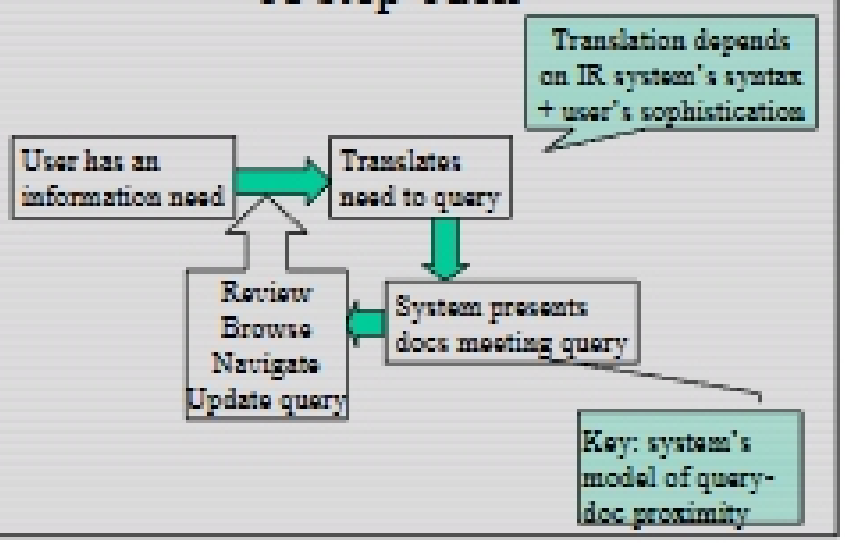
Lecture 5

April 23, 2001

Today's topics

- Generalized query operators
  - Evidence accumulation and structured queries
- Basics of Bayesian networks
  - Bayesian nets for Text Retrieval
- Structured+unstructured queries
  - Adding database-like queries

A step back



Models of query-doc proximity

- Boolean queries
  - Doc is either in or out for query
- Vector spaces
  - Doc has non-negative proximity to query
- Evidence accumulation
  - Combine score from multiple sources
- Bayesian nets and probabilistic methods
  - Infer probability that doc meets user's information need

Evidence accumulation

- View each term in the query as providing partial evidence of match
- *tf* × *idf* + vector space retrieval is one example
  - Corpus-*dependent* (*idf* depends on corpus)
- In some situations corpus-dependent evidence is undesirable

Corpus-independent evidence

- When is corpus-independent scoring useful?
  - When corpus statistics are hard to maintain
    - Distributed indices - more later
    - Rapidly changing corpora
  - When stable scores are desired
    - Users get used to issuing a search and seeing a doc with a score of (say) 0.9303
    - User subsequently filters by score
      - "Show me only docs with score at least 0.9"

Corpus-independent scoring

- Document routing is a key application
  - There is a list of standing queries
    - e.g., *bounced check* in a bank's email customer service department
  - Each incoming doc (email) scored against all standing queries
  - Routed to destination (customer specialist) based on best-scoring standing query
- More on this with automatic classification

Typical corpus-independent score

- Use a convex function of tf_{ij}
  - e.g., $\text{Score}(i,j) = 1 - \exp(-\alpha \times tf_{ij})$
  - α is a tuning constant
  - gives a contribution of query term i for doc j
- Given a multi-term query, compute the average contribution, over all query terms

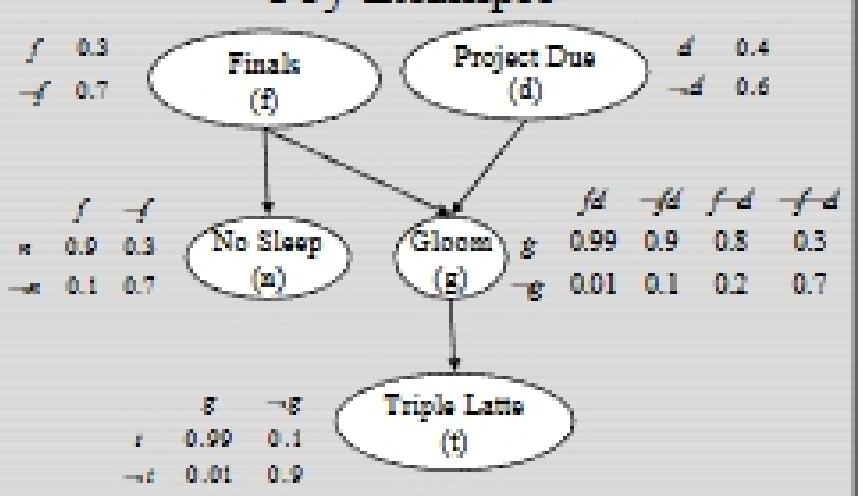
Bayesian Networks for Text Retrieval

- Text retrieval
  - Find the best set of documents that satisfies a user's information need
- Bayesian Network
  - Model causal relationship between events
  - Infer the belief that an event holds based on observations of other events

What is a Bayesian network?

- Is a directed acyclic graph
- Nodes
  - Events or Variables
    - Assume values.
    - For our purposes, all Boolean
- Links
  - model dependencies between nodes

Toy Example



Links as dependencies

- Link Matrix
  - Attached to each node
    - Give influences of parents on that node.
  - Nodes with no parent get a "prior probability"
    - e.g., *f*, *d*
  - interior node : conditional probability of all combinations of values of its parents
    - e.g., *s*, *g*, *t*