



**INLS 490-154: Information Retrieval Systems  
Design & Implementation  
Spring 2009**

**Wrap-up**

*Looking back, looking forward*

April 23, 2009

**Chirag Shah**

School of Information & Library Science (SILS)

UNC Chapel Hill

---



## Today's class

- Review of the course
- Pointers to some “advance” topics
- Discussion on related IR problems
- Course evaluation



# **COURSE REVIEW**



# Structured data access

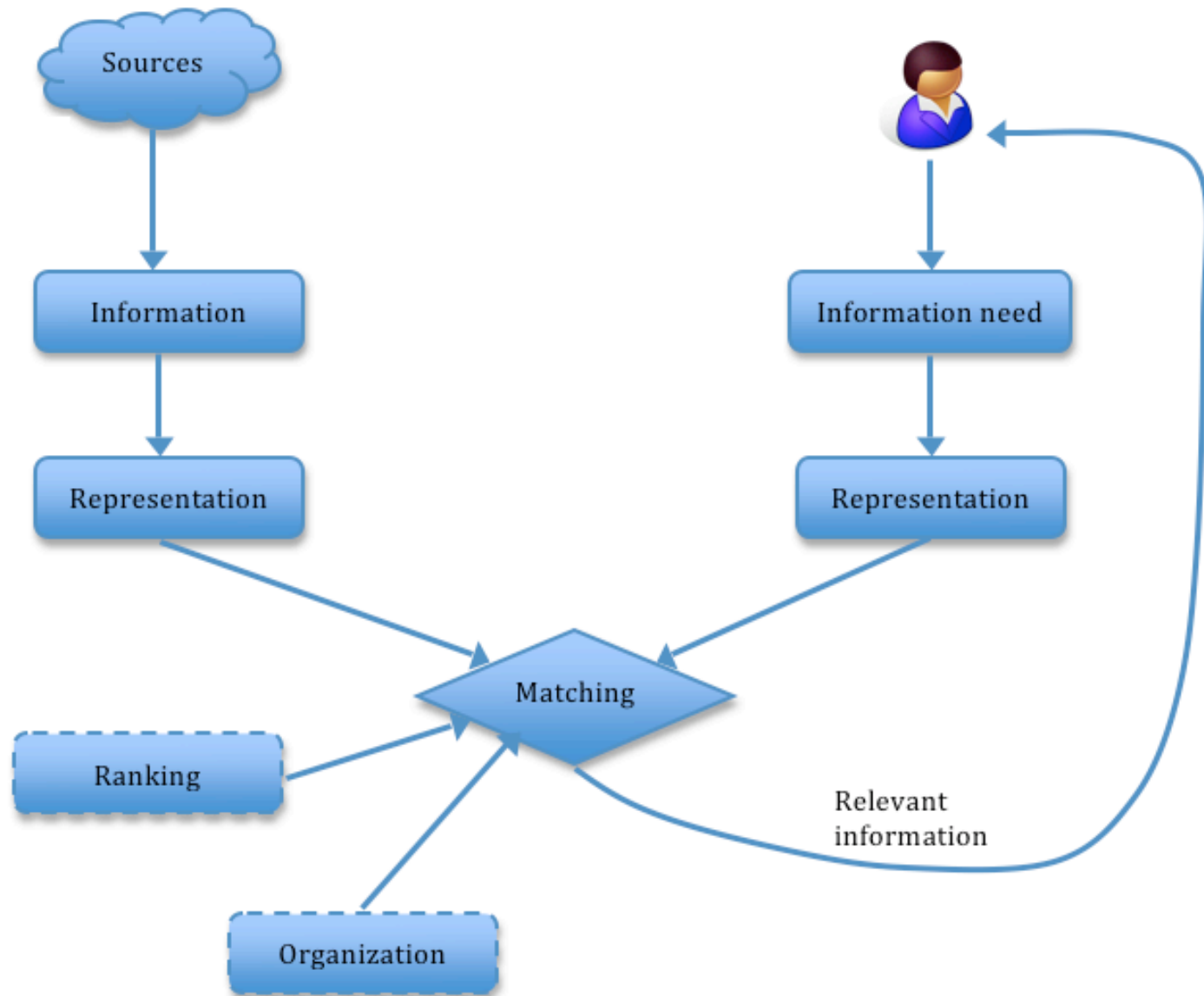
## Theory

- Storing structured information
- Accessing structured information
- Connecting backend with the front end
- Indexing
- Stopwords

## Practice

- MySQL database
- SQL
- PHP
- LIKE and MATCH expressions

# A model for Information Retrieval



# Lemur – this one and that one!



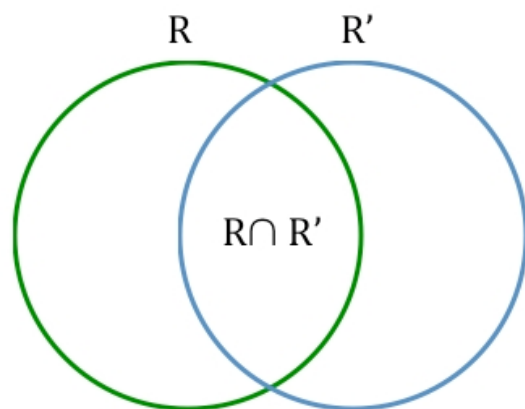
# Indexing

## Theory

- Tokenization
- Stopwords
- Stemming
- Storage
- Recall
- Precision

## Practice

- TREC format
- Porter's stemmer
- Krovetz's stemmer
- **ParseToFile**
- **BuildIndex**
- **dumpindex**



R= Relevant, R'=Retrieved

# Query processing and retrieval

## Theory

- Query representation
- Matching
- Cosine similarity
- Vector space model

## Practice

- ParseToFile
- BuildIndex
- RetEval
- TFIDF

# Retrieval models

## Theory

- Exact match
- Best match
- Language modeling
- Probabilistic models

## Practice

- Boolean (structured) queries
- **ParseInQueryOp**
- **StructQueryEval**
- **RetEval** with KL and Okapi

# Relevance feedback

## Theory

- “Relevance”
- Relevance feedback
- Pseudo-relevance feedback
- Query expansion

## Practice

- **RetEval** output fed into **RetFBEval**
- Presenting feedback terms to the user
- Incorporating user feedback to change the query or modify the term weights

# Evaluation

## Theory

- Measuring the query performance
- Measuring the system performance
- Comparing rank lists

## Practice

- `trec_eval`
- Recall
- Precision
- R-precision
- AP, MAP, GMAP
- bpref
- RR, MRR
- Pearson's covariance
- Spearman's Rho
- Kendall's Tau
- `R` statistical package

# UI for search

## Theory

- Design decisions for the backend
- Design decisions for the frontend

## Practice

- **indrid**
- **runquery**
- Changing '**runquery**' to print snippets in HTML
- **AJAX**



# IR on Web 2.0

## Theory

- Web 2.0
- REST architecture

## Practice

- API request using REST
- Parsing XML



# Web crawling

## Theory

- Data collection
- Crawling
- Harvesting

## Practice

- `wget`
- YouTube harvesting

# Information organization

## Theory

- Collection presentation
- Results organization
- k-means or centroid-based clustering
- Bisecting k-means or top-down clustering
- Agglomerative or bottom-up clustering

## Practice

- Term-clouds
- Cluster
- OfflineCluster



# **ADVANCE TOPICS**

## Latent Semantic Indexing (LSI)

- LSA applied to IR
- Reducing the dimensionality of the original term-document matrix using singular value decomposition (SVD)
- “Smart” SVD = success behind Google’s (and others) speed and success
- *S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. **Indexing by latent semantic indexing**. JASIS, 41(6): 391-407, 1990.*

## NLP for IR

- Use “better” terms for indexing
- Use NLP techniques to find related terms to be used for indexing
- Use NLP techniques to implement more sophisticated processing of queries
- *D. D. Lewis and K. Sparck Jones. **Natural language processing for information retrieval**. *Communications of the ACM*, 39(1):92-101, 1996.*

## Genomics IR

- Intersection of IR and bioinformatics
- Basic Local Alignment Search Tool (BLAST) – a heuristic approach
- Used to compare genomic DNA and protein sequences
- *S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman. **Basic local alignment search tool**. *Journal of Molecular Biology*, 215:403-410, 1990.*

# Image retrieval

- Two major approaches:
  1. Categorical or analytic keyword based searching
  2. Feature space searching
- *A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. IEEE PAMI, 2000.*



## Clustering

- Karger, Pedersen, and Tukey. *Scatter/Gather: a cluster-based approach to browsing large document collections*. In *Proceedings of ACM SIGIR 1992*. pp 318-329.
- O. Zamir and O. Etzioni. *Grouper: a dynamic clustering interface to web search results*. In *Proceedings of WWW Conference. 1999*.

## Etc.

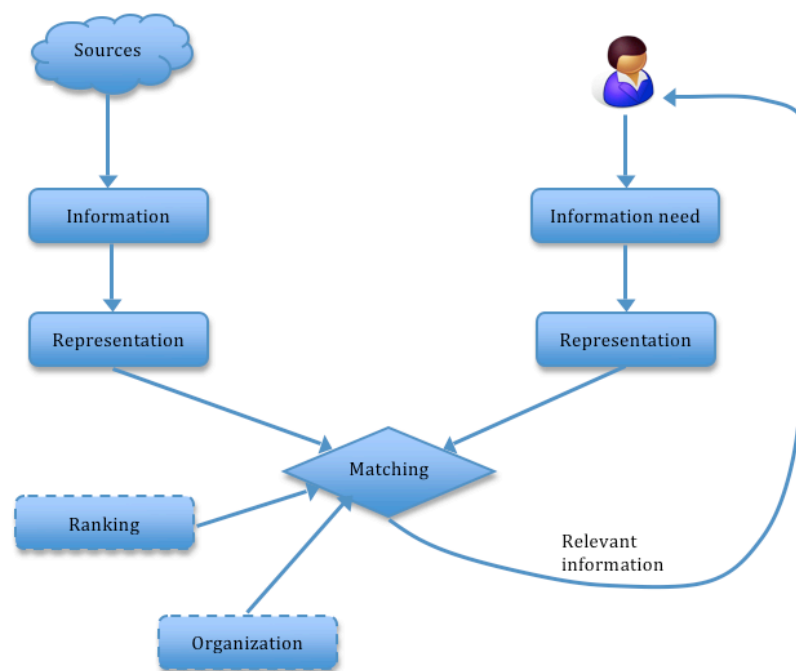
- J. Callan. *Distributed information retrieval*. In *Advances in Information Retrieval*, Kluwer Academic Publishers, 2000.
- V. Laverenko, M. Choquette, and W. B. Croft. *Cross-language relevance models*. SIGIR 2002.
- S. Dumais, E. Cutell, J. Cadiz, G. Jancke, R. Sarin, and D. Robbins. *Stuff I've seen: a system for personal information retrieval and re-use*. SIGIR 2003.



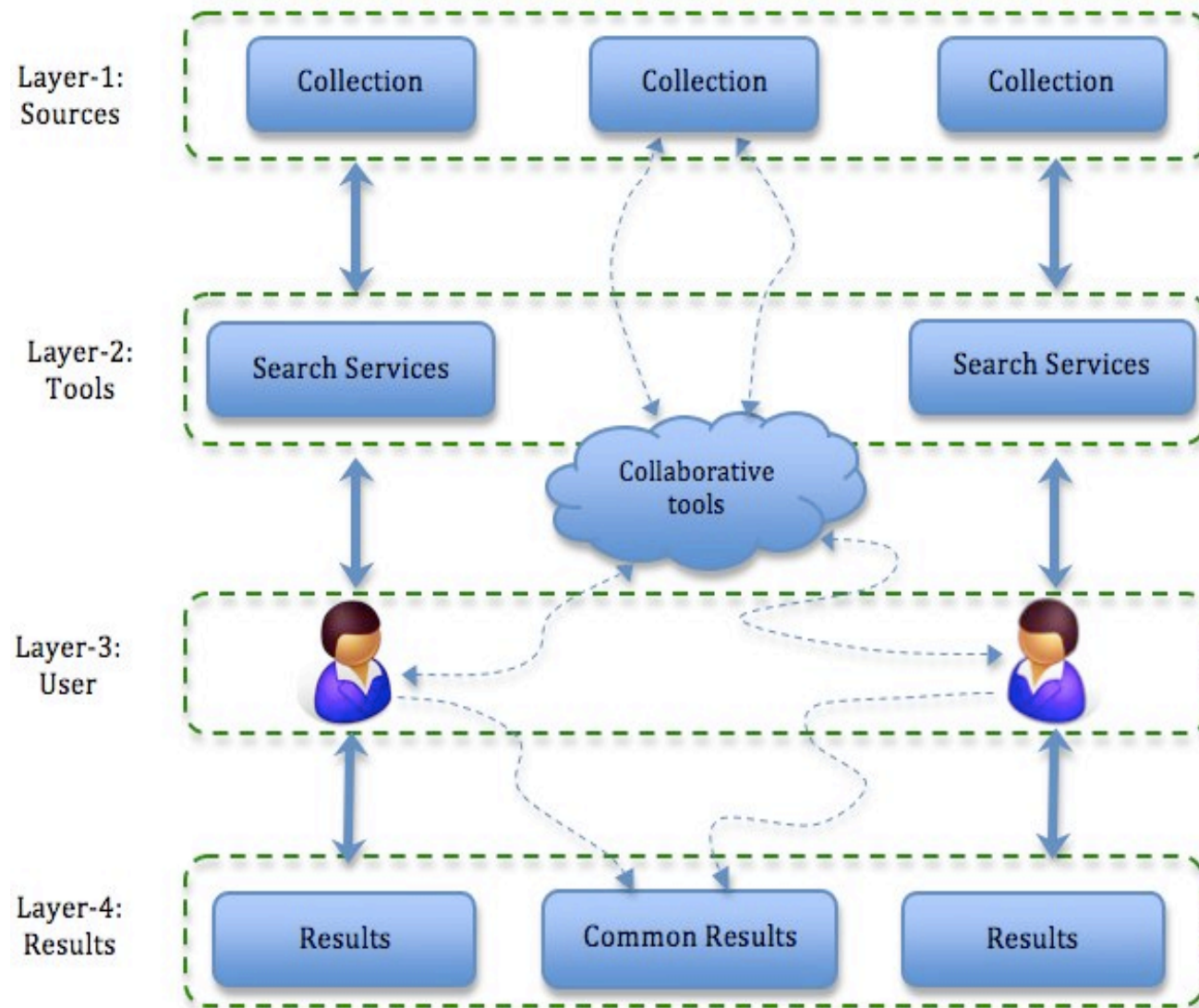
# **RELATED IR PROBLEMS**

# Topic Detection and Tracking (TDT)

- Story Link Detection (SLD)
- New Event Detection (NED)
- Tracking
- Hierarchical Topic Detection



# Collaborative IR





**That's all folks!**

**Thank you and good luck!**